RADC-TR-68-100
Final Report

ON-LINE INFORMATION RETRIEVAL USING ASSOCIATIVE INDEXING

Harold Borko
Donald A. Blankenship
Robert C. Burket

Systems Development Corporation

TECHNICAL REPORT NO. RADC-TR-68-100
May 1968

Rome Air Development Center
Air Force Systems Command
Griffiss Air Force Base, New York

ON-LINE INFORMATION RETRIEVAL USING ASSOCIATIVE INDEXING

Harold Borko
Donald A. Blankenship
Robert C. Burket
Systems Development Corporation

AFLC, GAFB, N.Y., 4 Jun 6[illegible]

FOREWORD

This document is the Final Report submitted by Systems Development Corporation, 2500 Colorado Avenue, Santa Monica, California, under Contract F30602-67-C-0077, Project 4594, Task 459401, for Rome Air Development Center, Griffiss Air Force Base, New York. SDC report number is TM-(L)-3851. Mr. Nicholas M. Di Fondi, EMIIH, was the RADC Project Engineer.

This technical report has been reviewed by the Foreign Disclosure Policy Office (EMLI) and the Office of Information (EMLS) and is releasable to the Clearinghouse for Federal Scientific and Technical Information.

This technical report has been reviewed and is approved.

Approved: Major James S. McGinnis
for FRANK J. COMAINI
Chief, Information Processing Branch
Intel & Info Processing Div.

Approved: Howard Dano
JAMES J. DIMEL, Colonel, USAF
Chief, Intel & Info Processing Div.

FOR THE COMMANDER: Irving J. Gabelman
IRVING J. GABELMAN
Chief, Advanced Studies Group

ABSTRACT

Experiments were performed to determine the feasibility of using ALCAPP as one form of on-line dialogue.

Assuming the ALCAPP (Automatic List Classification and Profile Production) system is in an on-line mode, investigations of those parameters which could affect its stability and reliability were conducted. Fifty-two full text documents were used to test how type of indexing, depth of indexing, the classification algorithm, the order of document presentation, and the homogeneity of the document collection would affect the hierarchical grouping programs of ALCAPP. Six hundred abstracts were used to study the effect on document clusters when more documents are added to the data base and the effect on the final cluster arrangement when the initial assignment of documents to clusters is arbitrary.

Results reveal that the only time significant differences in the classification of documents does not occur is when the order of document presentation is varied. Final clusters are significantly affected by the initial assignment of documents to clusters. The number of documents added to a data base allows stability of clusters only to a cutoff point which is some percentage of the original number of documents in the data base.

TABLE OF CONTENTS

TABLE OF CONTENTS--CONCLUDED

LIST OF FIGURES

LIST OF FIGURES--CONTINUED

LIST OF FIGURES--CONCLUDED

SECTION I

INTRODUCTION

Just as a library divides its collection of books into subject categories, so must an automated system organize its files into sections to achieve efficient storage and retrieval. However, if the classification of documents in an automated system is done manually, some of the advantages of the high-speed computer are lost, due to the delays in preparing the input. This problem has been recognized by researchers, and a number of attempts have been made to devise a classification algorithm that would be both reasonable and economically feasible.

In traditional classification systems, skilled librarians classify documents into categories on the basis of subject content. In an automated system, where the work of classification must be carried out by computers and not by people, class membership is determined on the basis of the words contained in the document or in a list of index terms ascribed to the document. This is a radically different principle, but it is a reasonable one. Ideas are expressed in words, and documents on different topics will use different sets of words to express ideas. It follows, therefore, that documents can be ordered into classes on the basis of similarity or differences in vocabulary. It is further postulated that classifying documents in accordance with the principle of similar word usage would result in a classification system analogous to, but not identical with, traditional subject categories and one that would be usable by both men and machines.

A number of mathematical techniques for deriving classification systems have been suggested. These include clump theory, factor analyses, latent-class analysis, discrimination analysis and others. References to these techniques along with a brief description may be found in Automated Language Processing (Borko, 1967). In general, all of the above procedures require lengthy computation and the amount of computer time increases by some factor, either the square or the cube, of data base size. As a result, these sophisticated taxonomic techniques are impractical when applied to large data bases.

Lauren Doyle (1966), in a research project supported by the Rome Air Development Center, devised a procedure for breaking this impasse, and he described a method of automatic classification that uses computer time in direct proportion--as a logorithmic function--to the number of items in the base. The programs, called ALCAPP (Automatic List Classification and Profile Production), are based upon the techniques of Joe Ward (Ward and Hook, 1963). Doyle's work was a major methodological contribution, for it removed a great obstacle from the path toward practical automatic document classification.

The current project was a continuation of the study of automatic classification techniques and had as its major tasks:

(1) To investigate the statistical reliabilities of the ALCAPP algorithms.

(2) To evaluate the effectiveness of the machine-produced classification hierarchy as an aid in predicting document content and as an adjunctive retrieval tool in an on-line time-shared system.

(3) To recode the ALCAPP programs for operation on the GE 635 computer which is available for use at the Rome Air Development Center.

SECTION II

SELECTION OF THE DATA BASE

Since the RADC contract, under whose sponsorship these studies were conducted, did not specify the subject content of the data base, it was decided to use documents in the field of information science. The main advantages are that these documents are readily available at System Development Corporation and that SDC employs a number of experts in this area. If necessary, these people could be used to evaluate the reasonableness of the data processing results, e.g., indexing and classification, and the effectiveness of the system. On the negative side, information science does not have a well-specified thesaurus or authority list of terms for use in indexing.

After consultation with the contract monitors at RADC, it was agreed that the advantages of using a data base of information science materials outweighed the disadvantages. With their concurrence, the following documents were selected:

(1) The full text of the 52 papers that were printed in the Proceedings of the 1966 American Documentation Institute Annual Meeting (Black, 1966).

(2) The abstracts of 600 other documents in the field of information science.

Simplified keypunching rules were specified by the contractor and approved by the monitor (see the Appendix). The entire data base--that is, both the 52 documents and the 600 abstracts--was keypunched in accordance with these rules and was thus made available for computer processing. The 52 full-text documents were used to study the reliability and consistency of the automatic classification procedure. A subset of these documents was used in the experiments judging the incremental value of the classification hierarchy in predicting document content and relevance. The abstracts were used to study the stability of the classification categories as new documents are added to the data base.

SECTION III

PREPARATION OF WORD LISTS

The data used as input to the classification programs were lists of index terms derived from the documents, and not the documents themselves or their natural language abstracts. By indexing each document both manually and by machine-aided methods, the type and quality of the indexing was varied. The length of the word lists was also varied by creating lists of 6, 15, and 30 terms each. Thus, it was possible to determine the effect that the type and depth of indexing would have on the reliability and consistency of the resulting classification systems.

1. PREPARING WORD LISTS FROM THE 52 FULL-TEXT DOCUMENTS

The 52 full-text documents were to be used to investigate the effect of indexing type and indexing depth on the reliability and consistency of the ALCAPP classification procedures. To do so, each document was indexed by six different procedures as follows:

Human Indexing	30 terms
Human Indexing	15 terms
Human Indexing	6 terms
Machine-Aided Indexing	30 terms
Machine-Aided Indexing	15 terms
Machine-Aided Indexing	6 terms

a. Human Indexing

The "human indexing" was done by trained librarians from the SDC library staff. They were given copies of the 52 documents and asked to assign 30 appropriate subject headings. They were asked to use a free vocabulary, since no authority list was available. They were also instructed to arrange the terms in a rough order of importance, so that, for each document, the first 6 terms, the first 15 terms, and the complete list of 30 terms could be used separately for different phases of the experiment. In some instances, the indexers found it impossible to list 30 terms, and shorter lists were accepted.

Since there were no controls over the vocabulary, some editing was necessary in order to achieve a degree of consistency and compatibility. The index terms were keypunched, and sorted alphabetically, by frequency, and by individual document. These lists were returned to the indexers for editing and modification. Variations in the use of plural and singular endings were changed, e.g., COMPUTER was changed to COMPUTERS; certain modifiers were dropped, e.g., MAGNETIC TAPE STORAGE was changed to MAGNETIC TAPE; word order was standardized, e.g, ABSTRACTING, AUTOMATIC become AUTOMATIC ABSTRACTING; near synonyms were combined, e.g., AUTHORITY LISTS was merged into AUTHORITY FILES; and some names were abbreviated, e.g., COMMITTEE ON SCIENTIFIC AND TECHNICAL INFORMATION became COSATI, etc.

The sole aim of the editing was to achieve consistency in the use of terms for this experiment. It was not our purpose to create a generally useful lexicon. No attempt was made to combine generally similar terms

into a single concept if the indexer believed them to be separate, so that ALGEBRA, ABSTRACT and ALGEBRA, MODERN were retained as separate terms. Similarly, if a single document was indexed by both COMMUNICATION and COMMUNICATION OF TECHNICAL INFORMATION, both terms were retained.

For mechanical reasons and in order to reduce computer processing time, each term was truncated at 15 alpha characters. In those instances where truncation could cause ambiguity, the numeric digits, 1, 2, 3, etc., were added to insure uniqueness.

b. Machine-Aided Indexing

While it was not the purpose of this study to devise methods of automatically indexing textual material, the project staff did process the documents in the data base and prepared word lists as aids in the selection of index terms. Each of the 52 complete documents was processed individually to create an alphabetical list of all words used in the text, together with their frequency of occurrence. This basic list was then reordered so that the word with the highest frequency would be listed first and the others would follow in descending order. Then, using an available routine that would combine plural and singular forms of the same root, the alphabetically ordered list was rerun and words with the same root combined. Next the individual lists of all 52 documents were merged, creating a unified frequency-ordered list in which singulars and plurals were combined. An alphabetically ordered listing was also obtained.

MANUAL INDEXING

DOCUMENT NO. 1: Progress in Internal Indexing[1]

---

information storage and retrieval systems
indexing
automatic indexing
indexing, manual
abstracting and indexing services
subject indexing
computers--applications
computers--applications--writing and editing
content analysis (computers)
machine translation
cataloging of technical literature
computers--machine-readable text
computers--research
information science--research
cataloging
documentation
data processing systems--libraries
computers--applications--libraries
report writing
research--indexes
congresses and conventions--abstracting and indexes
books--abstracting and indexes
word files
sentence files
punched cards
sentence entities
recursive procedures
indexing term selection
term dictionaries
internal indexing

Figure 1. Manual Indexing of Document No. 1 before Truncation

---

[1] Maloney, C. J. and M. H. Epstein. Progress in Internal Indexing. (Black, 1966)

MACHINE-AIDED INDEXING

DOCUMENT NO. 1: Progress in Internal Indexing[2]

---

term	machine
index	list
word	core
dictionary	system
sentence	memory
table	internal
text	bits
output	user
computer	external
file	context
character	cards
report	purged
input	publication
tape	format
program	coordinate

Figure 2. Machine-Aided Indexing of Document No. 1

---

[2]Maloney, C. J. and M. H. Epstein. Progress in Internal Indexing. (Mack, 1966)

MACHINE INDEXING: ABSTRACT

Title: Identifying and Locating Standards

---

1. standard
2. subject
3. number
4. type
5. report
6. association
7. national
8. organization
9. area
10. international
11. requirement
12. individual
13. code
14. identify
15. librarian
16. quality
17. symbol
18. LSCA (abstractor's code)
19. identification
20. deal
21. produced
22. difficulty
23. microfiche
24. image
25. cover
26. practice
27. initial
28. firm
29. encountered
30. specification

Figure 3. Example of Machine-Aided Indexing of an Abstract

The task of the editor was to select 30 words for each of the 52 documents for input to the Hierarchical Grouping Program. Each list had to be so arranged that the first 6 terms and the first 15 terms could themselves constitute lists for processing. Using these various computer prepared printouts, the editor was able to make the selection reasonably efficiently, and to prepare the lists for subsequent computer processing.

2. PREPARING WORD LISTS FROM THE 600 DOCUMENT ABSTRACTS

The 600 document abstracts were used in the experiments designed to test the stability of the groupings which result from the application of the Cluster Finding Program. Word lists had to be prepared from each of the abstracts for input to this program. These lists were prepared by computer analysis on the basis of frequency of word occurrences and then minimally edited by the investigators. The abstracts were not indexed manually, since our objective was not to determine whether there would be differences in the clustering due to differences between manual and machine indexing. The extent of such differences would be determined more precisely in the hierarchical structuring experiments, using full documents. With the 600 abstracts, our only purpose was to measure the stability of the resulting clusters, and for this purpose we used lists of 30 terms prepared by computer analysis of the abstracts.

SECTION IV

MEASURING THE RELIABILITY AND CONSISTENCY OF THE ALCAPP SYSTEM

A classification system is considered to be reliable if documents classified into a given category will be classified into that same category on subsequent trials. If the system is not reliable and the document classifications vary, classification will not be a useful adjunct to retrieval. Reliability and consistency are necessary, but not sufficient, conditions for a useful classification system. Therefore, the first series of experiments were designed to investigate the reliability and consistency of the ALCAPP system and the variables that affect the reliability of the automatic classifications.

1. DESCRIPTION OF THE HIERARCHICAL GROUPING PROGRAM

It is obvious that different classification procedures will result in different classification structures. The Library of Congress Classification schedule differs from the Dewey Decimal, and clearly a machine-derived system will differ from both of these. The task of this project was to determine the statistical properties of the ALCAPP method of machine classification, and not compare it with manual methods.

Machine classification is based on the assumption that documents containing more words in common are more similar to each other in content than are documents which have fewer words in common. Each document is represented by a surrogate or list of index terms. As was pointed out in the previous section, these index terms could be derived either manually or by machine. The ALCAPP programs begin by comparing these lists and

counting the number of identical terms in each pair. It constructs a matrix--or rather a half matrix, since the data are symmetric around the diagonal--in which each cell contains the number of terms the two word lists share in common. In this study, 52 document index lists were compared, so the actual number of comparisons is

$$\frac{52 \times 51}{2} = 1,326.$$

The program searches the matrix and finds the largest cell value, i.e., that pair of documents with the most terms in common. In case of tie, the first value is chosen. These two documents--let's call them D_i and D_j--are now chosen to be the first pair in the hierarchical classification structure. This completes the first iteration of the program.

In the second iteration, documents i and j have been eliminated and combined into one value--call it G_1. A new matrix is created of order N-1, or 51. Documents i and j are excluded, but in their place is a new vector G_1. The program now calculates the similarity of the remaining documents to G_1 and places this value in the appropriate cell. It then searches the matrix for the highest value. This value can represent the similarity of two documents, as was the case in the first iteration, or it can represent the similarity of a document with G_1. If the former is true, the two documents will be combined to form a new group of two, while in the latter instance, a third document will

be added to the first group. In either case, the new group is called G_2, and the program has completed the second iteration.

To complete the entire hierarchical grouping structure, one less iteration than there are documents in the set will be required. The last iteration will form a single group containing the entire collection.

A mathematical description of the classification process can be found in the documents by Ward (1959), Ward and Hook (1963), and Baker (1965). By using the same basic technique but varying the function used to calculate similarity, different hierarchical structures can be formed. (Figures 4 and 6 are examples.) The program can also label the nodes of the structure, thus providing an indication of the common elements that link the documents together (see Figures 5 and 7 as examples).

2. MEASURING CLASSIFICATION SIMILARITY

Classification similarity is measured by means of a distance matrix that provides a measure of the distance separating each document from every other document in the classification system. This procedure is used to provide a rigorous measure of the reliability and consistency of automatic classification under different laboratory conditions. For these purposes a small, intensively analyzed sample of 52 documents was considered.

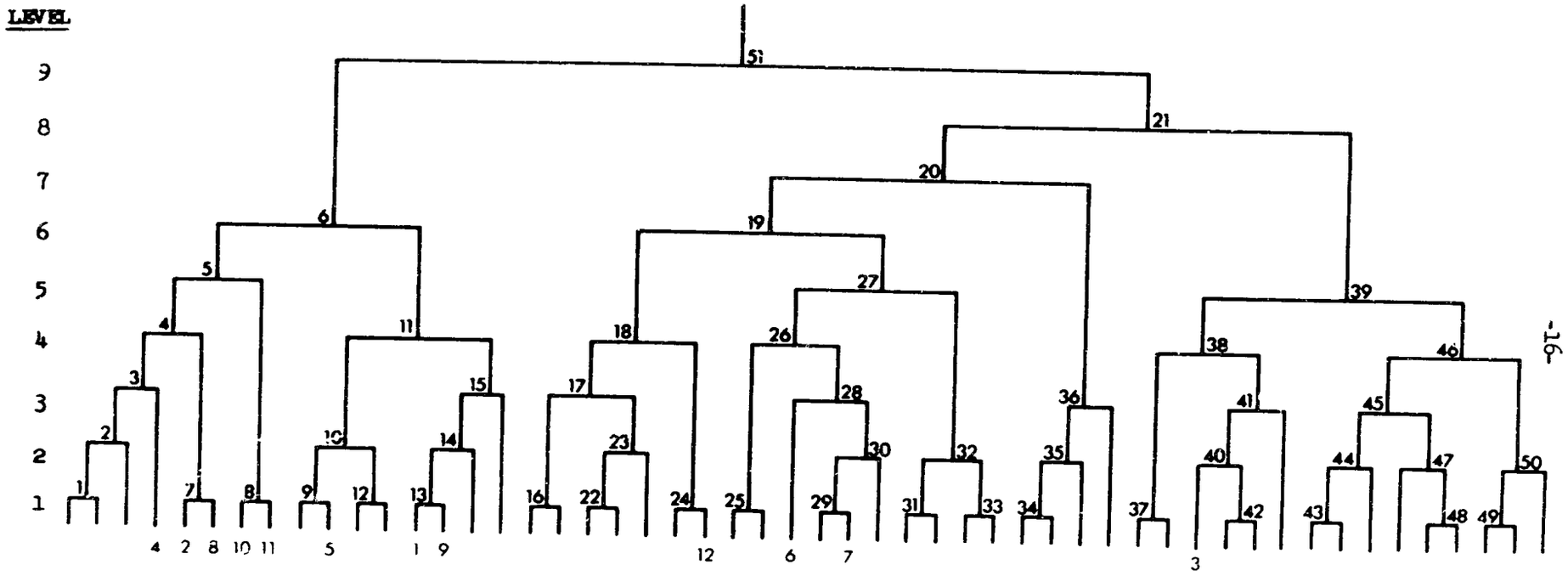

Figure 4. Hierarchical Clustering Scheme 2, Group D2

HIERARCHICAL CLUSTERING SCHEME 2

1	literature reference library subject search	14	tape page publication sentence code	27	question research concept test analysis	40	sheet subsystem interest profile file
2	group literature paper reference form	15	code text character font input	28	terms chemical evaluation experiment level	41	interest notification SDI sheet statistics
3	facet group classification thesaurus literature	16	storage card request file question	29	experiment terms abstract question search	42	sheet subsystem interest library data
4	classification group automatic coordinate documentation	17	request example figure requestor synonym	30	terms chemical experiment answer concept	43	organization scientific vocabulary need technical
5	classification thesaurus category EJC group	18	request record card storage book	31	model distribution figure question analysis	44	scientific need organization vocabulary science
6	word program time subject list	19	question search request research subject	32	distribution factor model set significant	45	need scientific center service dissemination
7	similarity matrix automatic classification data	20	question search retrieval information computer	33	set significant variable method result	46	need scientific service technical user
8	descriptor EJC section thesaurus technical	21	information system document index user	34	image microfiche access microfilm file	47	center need problem result service
9	dictionary language word term retrieval	22	language request subject search term	35	image microfilm frame keyboard microfiche	48	center service need technical abstract
10	dictionary citation rule tool language	23	example synonym figure relationship value	36	microfilm film frame image microfiche	49	facility paper language file data
11	dictionary tape list code format	24	catalog item material machine record	37	article keyword content journal title	50	education facility value language paper
12	citation literature title form list	25	engineer concept science study project	38	interest profile subsystem category article	51	information system index document retrieval
13	publication sentence text format tape	26	concept terms research answer chemical	39	user data library abstract interest		

Figure 5. Hierarchical Clustering Scheme 2

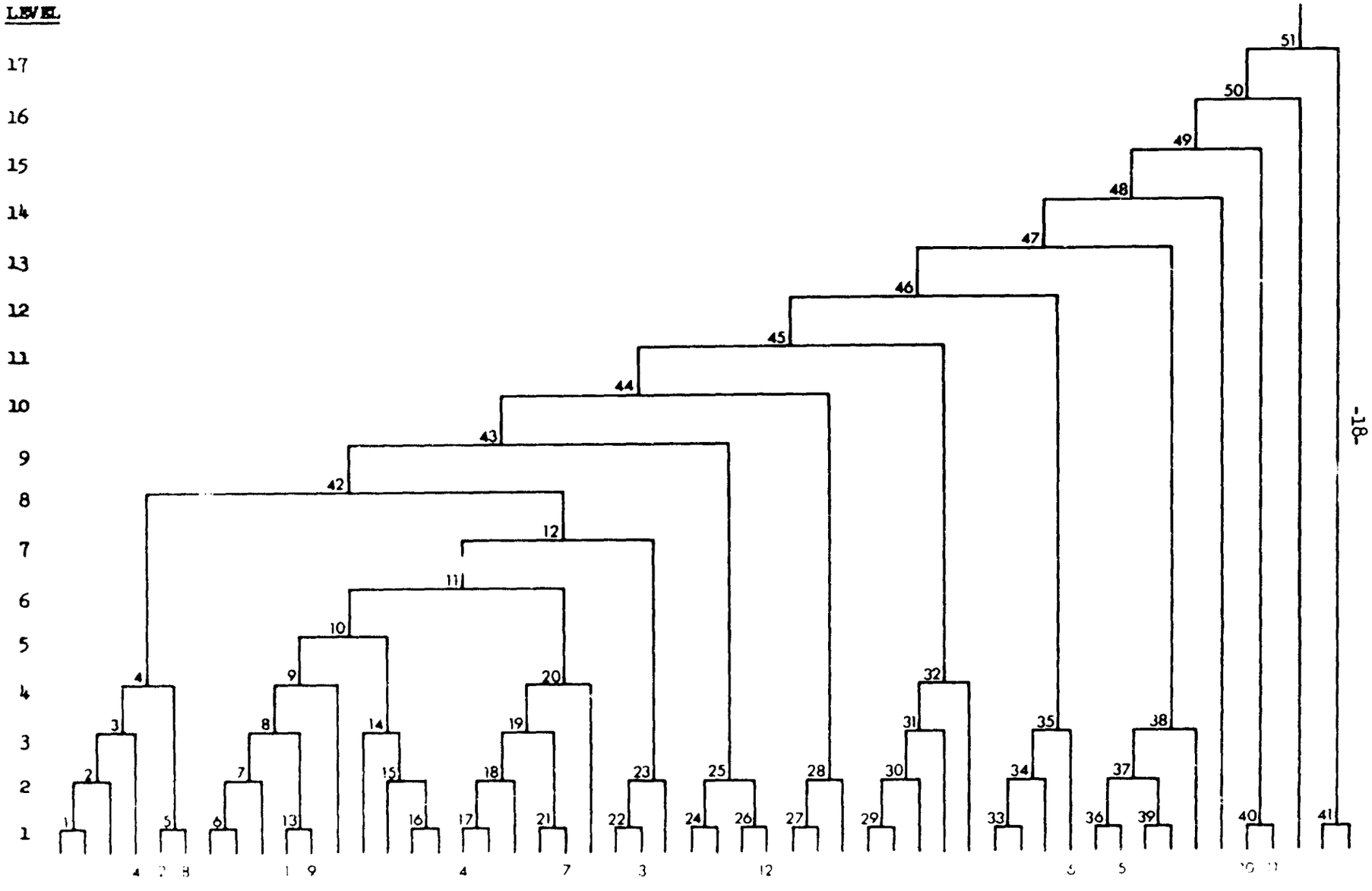

Figure 6. Hierarchical Clustering Scheme 3, Group E2

HIERARCHICAL CLUSTERING SCHEME 3

1	literature reference library subject search	14	need scientific service center organization	27	facility paper language file data	40	descriptor FIC section thesaurus technical
2	group literature paper reference form	15	center service need scientific machine	28	facility paper author language form	41	model distribution figure question analysis
3	facet group classification thesaurus literature	16	center service need technical abstract	29	image microfiche access microfilm file	42	system index search retrieval information
4	classification group automatic coordinate documentation	17	language request subject search term	30	image microfilm frame keyboard microfiche	43	system information index search document
5	similarity matrix automatic classification date	18	chemical answer problem request language	31	microfilm film frame image microfiche	44	system information index document retrieval
6	storage card request file question	19	answer chemical terms experiment problem	32	microfilm page film frame image	45	system information index document retrieval
7	mechanized tape card request storage	20	answer chemical experiment relationship terms	33	engineer concept science study project	46	system information index document retrieval
8	tape format card coordinate IBM	21	experiment terms abstract question search	34	engineer study table concept science	47	information system index document retrieval
9	tape file author format card	22	article keyword content journal title	35	engineer study field research table	48	information system index document retrieval
10	file card need scientific tape	23	article keyword subsystem content journal	36	dictionary language word term retrieval	49	information system index document retrieval
11	search question retrieval user computer	24	read interest report subject document	37	dictionary citation rule tool language	50	information system index document retrieval
12	search user retrieval index system	25	item material catalog department flow	38	tool dictionary language citation education	51	information system index document retrieval
	publication sentence text format flow	26	catalog item material machine term	39	citation literature title term list		

Figure 7. Hierarchical Clustering Scheme 3

Document No.

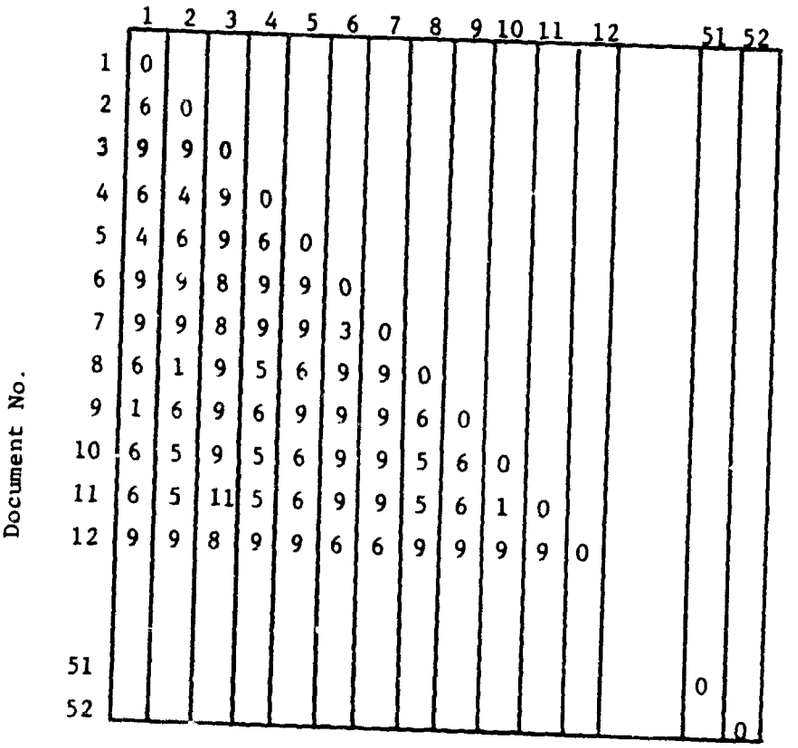

Figure 8. Distance Matrix Based upon Figure 4

In these experiments, the distance matrix was a symmetrical matrix of order 52, for we were using 52 documents. The number in each cell (the intersection of each row and column) represents the number of the level at which the two documents are joined. A distance matrix was computed for the 12 documents distributed on the hierarchical clustering scheme, as illustrated in Figure 4. The half-matrix of distances is shown in Figure 8.

Once the distance matrices have been computed, it now becomes possible to determine the degree of similarity between any or every two matrices by correlating the respective columns. Thus, it also becomes possible to measure the importance that such variables as the depth or type of indexing would have on the similarity of the resulting classifications.

3. VARIABLES RELATED TO CLASSIFICATION SIMILARITY

What makes one classification scheme similar to another? What variables affect the degree of similarity between two classification schemes? These studies were designed to shed some light, in the form of statistical data, on the intuitive answers that are usually given to the above questions.

Based upon a logical analysis, the following five variables are believed to be related to classification similarity:

(1) The homogeneity of the document collections.

(2) The classification procedure, or algorithm, being used.

(3) The type or quality of indexing--whether it be term or concept indexing.

(4) The depth of indexing used.

(5) The order in which the documents are processed.

These variables were compared systematically in order to determine their effect on the resulting classification structures.

a. The Homogeneity of the Document Collection

One of the variables that could affect the reliability of the classification procedure is the homogeneity or diversity of the document collection. To test the effect of this variable, we would need four or five different document collections, and these collections would have to span a range from a narrow hard science collection, such as solid state physics, through perhaps the broader field of geology on to the still broader field of social sciences. This project is basically a pilot study, and because of time and cost constraints, we decided not to manipulate the data base as a variable in this experimental design. Instead, we kept the homogeneity factor constant by limiting the analysis to the field of information science documentation. The selected collection of documents probably constitutes a mid-range position on the scales of diversity and hardness of data, for it covers a single, relatively homogeneous but broad subject area in the social sciences.

b. The Classification Algorithm

In the course of SDC's research program on automated classification, a number of different algorithms have been developed. While all of them use basically the same technique described in Paragraph 1, they do differ in the averaging function used--the mathematical formulas for computing a value of group similarity.

Two such algorithms seemed particularly worth investigating and comparing. These were arbitrarily called WD-2 and WD-3 in a sequence of modifications. The WD-2 algorithm maximizes the within-group similarity function and puts a premium on preserving the homogeneity of groups that have already been formed. The WD-3 algorithm takes an opposite approach and combines lists that have a minimum dissimilarity as contrasted with a maximum similarity.

While it might appear on the surface that these functions should perform similarly in forming groups, this need not be the case, for the program examines different data (see Figures 4 and 5). By including both programs in the experimental design, it was possible to compare the form and reliabilities of the classification structures.

c. The Type (or Quality) of Indexing

The documents to be classified were indexed by qualified librarians who were instructed to use multi-word concept or subject indexing. In addition, these same documents were indexed by key words using machine-aided selection techniques. Obviously, the lists differ.

The question being investigated is: Do classification systems based on human indexing differ significantly from classification systems based upon machine-aided indexing?

d. The Depth of Indexing

Since the inputs to the classification program are lists of words, it was important to investigate the effect that different-sized lists would have on the reliability of the classification structure. In order to test the effect of this variable, different length lists, containing 6, 15, and 30 terms each, were used and varied systematically.

e. The Order of Document Presentation

In the description of the hierarchical grouping program (Paragraph 1), it was explained that, although the program combined documents into groups by searching the similarity matrix for the highest cell value, when more than one cell had the same value, the first position was used to form the group. As a result of the procedure used, the order in which the documents are processed could affect the final hierarchical classification structure.

A series of experiments were designed to determine whether the order of document presentation would cause significant differences. The 52 documents were arranged in three different orders for input to the computer program. The documents were numbered from 1 to 52. The first order arranged the documents in a ascending numerical value. The second order was the reverse, with document number 52 being

processed first. And the third order was a random arrangement of the documents. For each of these three arrangements, hierarchical groupings of the 52 documents were computed and their structures compared for similarity.

4. THE EXPERIMENTAL DESIGN

The aim of this set of experiments was to investigate the reliability and consistency of automatically derived classification hierarchies, as selected attributes are varied in a controlled fashion. The four selected attributes are:

(1) The classification algorithm:

WD-2

WD-3

(2) The type of document indexing:

M = machine-aided

H = human

(3) The depth of indexing:

6 terms

15 terms

30 terms

(4) The order of document input for processing:

01 = ascending order 1-52

02 = descending order 52-1

03 = random order

In order to vary the attributes systematically, under all possible conditions, 36 hierarchical classification structures were required (2x2x3x3). Figure 9 lists all 36 classification matrices and the particular attributes that were used in their construction.

Once the classification structures were derived by machine processing, the information contained therein was transformed into sets of distance matrices, which were, themselves, correlated. The outcome of the correlation program was a 36 x 36 matrix, in which the rows and columns are the 36 different classification structures, and a cell value is the correlation coefficient indicating the similarity between the pair of classification schemes. The complete correlation matrix is reproduced as Figure 10.

The following criteria were used in interpreting the correlation matrix:

(1) High Similarity	r = .70 to .99
(2) Moderate Similarity	.40 to .69
(3) Slight Similarity	.20 to .39
(4) No Similarity	.00 to .19

These numbers and ranges are useful in making comparative judgments and not for absolute scalar judgments.

	Algorithm	Type	Depth	Order		Algorithm	Type	Depth	Order
1	WD-2	M	6	01	19	WD-2	H	6	01
2	WD-2	M	15	01	20	WD-2	H	15	01
3	WD-2	M	30	01	21	WD-2	H	30	01
4	WD-3	M	6	01	22	WD-3	H	6	01
5	WD-3	M	15	01	23	WD-3	H	15	01
6	WD-3	M	30	01	24	WD-3	H	30	01
7	WD-2	M	6	02	25	WD-2	H	6	02
8	WD-2	M	15	02	26	WD-2	H	15	02
9	WD-2	M	30	02	27	WD-2	H	30	02
10	WD-2	M	6	03	28	WD-3	H	6	02
11	WD-2	M	15	03	29	WD-3	H	15	02
12	WD-2	M	30	03	30	WD-3	H	30	02
13	WD-3	M	6	02	31	WD-2	H	6	03
14	WD-3	M	15	02	32	WD-2	H	15	03
15	WD-3	M	30	02	33	WD-2	H	30	03
16	WD-3	M	6	03	34	WD-3	H	6	03
17	WD-3	M	15	03	35	WD-3	H	15	03
18	WD-3	M	30	03	36	WD-3	H	30	03

Figure 9. Classification Structures Showing Systematic Variation of Attributes

Array	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
1	100	18	20	28	16	4	84	24	18	78	)	21	37	19	3	35	16	2	-1	
2	18	100	21	15	46	14	23	58	28	19	53	24	21	43	16	16	43	15	2	
3	20	21	100	11	20	26	21	24	74	22	32	57	17	28	23	13	22	24	6	
4	28	15	11	100	24	31	38	22	12	48	21	19	86	26	30	96	27	27	1	
5	16	46	20	24	100	55	18	41	32	15	54	33	26	84	54	28	85	55	3	
6	4	14	26	31	55	100	3	17	25	6	24	28	27	55	87	33	45	91	2	
7	84	23	21	38	18	3	100	30	20	90	26	23	51	22	3	45	21	1	3	
8	24	58	24	22	41	17	30	100	31	31	48	25	34	39	18	24	35	15	2	
9	18	28	74	12	32	25	20	31	100	20	49	61	19	39	29	13	38	28	5	
10	78	19	22	48	15	6	90	31	20	100	23	26	60	21	7	53	19	5	1	
11	20	53	32	21	54	24	26	48	49	23	100	52	30	61	27	22	63	25	5	
12	21	24	57	19	33	28	23	25	61	26	52	100	25	38	35	21	37	34	8	1
13	37	21	17	86	26	27	51	34	19	60	30	25	100	31	27	87	31	24	2	
14	19	43	28	26	84	55	22	39	39	21	61	38	31	100	53	30	89	54	5	
15	3	16	23	30	54	87	3	18	29	7	27	35	27	53	100	31	44	95	2	
16	35	16	13	96	28	33	45	24	13	53	22	21	87	30	31	100	29	29	2	
17	16	43	22	27	85	45	21	35	38	19	63	37	31	89	44	29	100	46	4	
18	2	15	24	27	55	91	1	15	28	5	25	34	24	54	95	29	46	100	3	
19	-1	2	6	1	3	2	3	2	5	1	5	8	2	5	2	2	4	3	100	3
20	8	-1	5	2	2	0	9	-2	5	6	6	12	5	2	1	2	2	1	30	10
21	2	1	1	4	2	3	0	1	1	-2	1	-2	2	4	2	4	4	4	5	1
22	-7	-2	-1	19	2	-1	2	1	2	5	1	-2	17	2	-2	15	6	-1	33	2
23	-6	-3	-1	17	3	-2	3	-1	1	5	1	-1	15	5	0	14	7	0	30	3
24	3	-1	-4	4	7	3	0	-1	-2	0	3	-4	4	8	1	4	7	2	6	1
25	2	-4	0	1	-3	-3	3	-4	1	2	-3	0	0	-3	-3	1	-4	-3	64	3
26	7	-2	2	-1	-3	-2	8	-3	3	6	1	5	2	-3	-1	0	-3	-1	30	6
27	2	1	0	5	5	4	0	0	0	-2	2	0	3	7	3	5	6	6	6	1
28	-9	-2	0	22	3	1	1	1	2	5	1	-3	19	3	0	18	7	1	33	2
29	-8	-2	-2	20	5	0	-1	0	1	1	2	-1	16	6	1	17	9	1	27	3
30	3	-1	-4	4	7	3	0	-2	-1	0	4	-3	4	8	2	4	7	2	7	1
31	7	-3	3	2	0	-2	7	0	2	7	0	1	1	-1	-3	2	-2	-2	60	2
32	7	-1	4	5	1	0	10	-3	5	7	5	11	7	1	0	5	1	0	36	8
33	2	1	-1	4	5	4	-1	1	-1	-2	2	-1	2	7	3	4	6	6	6	1
34	-7	-3	-1	20	0	-2	2	1	1	6	0	-3	18	0	-2	16	4	-2	31	1
35	-8	-1	-1	15	9	1	-1	0	2	1	3	-1	12	11	2	13	14	2	26	3
36	3	-1	-4	4	7	3	0	-2	-1	0	4	-3	4	8	2	4	7	2	7	1

Figure 10. Intercorrelation Matrix of 36 Classifica

A

17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
16	2	-1	8	2	-7	-6	3	2	7	2	-9	-8	3	7	7	2	-7	-8	3
43	15	2	-1	1	-2	-3	-1	-4	-2	1	-2	-2	-1	-3	-1	1	-3	-1	-1
22	24	6	5	1	-1	-1	-4	0	2	0	0	-2	-4	3	4	-1	-1	-1	-4
27	27	1	2	4	19	17	4	1	-1	5	22	20	4	2	5	4	20	15	4
85	55	3	2	2	2	3	7	-3	-3	5	3	5	7	0	1	5	0	9	7
45	91	2	0	3	-1	-2	3	-3	-2	4	1	0	3	-2	0	4	-2	1	3
21	1	3	9	0	2	3	0	3	8	0	1	-1	0	7	10	-1	2	-1	0
35	15	2	-2	1	1	-1	-1	-4	-3	0	1	0	-2	0	-3	1	1	0	-2
38	28	5	5	1	2	1	-2	1	3	0	2	1	-1	2	5	1	1	2	-1
19	5	1	6	-2	5	5	0	2	6	-2	5	1	0	7	7	-2	6	1	0
63	25	5	6	1	1	1	3	-3	1	2	1	2	4	0	5	2	0	3	4
37	34	8	12	-2	-2	-1	-4	0	5	0	-3	-1	-3	1	11	-1	-3	-1	-3
31	24	2	5	2	17	15	4	0	2	3	19	16	4	1	7	2	18	12	4
89	54	5	2	4	2	5	8	-3	-3	7	3	6	8	-1	1	7	0	11	8
44	95	2	1	2	-2	0	1	-3	-1	3	0	1	2	-3	0	3	-2	2	2
29	29	2	2	4	15	14	4	1	0	5	18	17	4	2	5	4	16	13	4
100	46	4	2	4	6	7	7	-4	-3	6	7	9	7	-2	1	6	4	14	7
46	100	3	1	4	-1	0	2	-3	-1	6	1	1	2	-2	0	6	-2	2	2
4	3	100	30	5	33	30	6	64	30	6	33	27	7	60	36	6	31	26	7
2	1	30	100	11	20	39	11	34	62	10	20	31	11	29	86	11	19	34	11
4	4	5	11	100	6	10	36	7	13	91	7	11	36	6	10	89	7	13	36
6	-1	33	20	6	100	69	24	35	15	7	95	63	24	39	21	7	98	60	24
7	0	30	39	10	69	100	33	33	32	12	71	83	33	26	39	13	70	85	33
7	2	6	11	36	24	33	100	4	11	34	26	30	100	4	7	35	25	34	100
-4	-3	64	34	7	35	33	4	100	35	5	34	23	4	75	38	6	34	21	4
-3	-1	30	62	13	15	32	11	35	100	12	14	28	12	26	64	12	15	27	12
6	6	6	10	91	7	12	34	5	12	100	8	13	35	5	8	97	7	14	35
7	1	33	20	7	95	71	26	34	14	8	100	65	26	37	21	8	95	63	26
9	1	27	31	11	63	83	30	23	28	13	65	100	30	17	31	13	65	88	30
7	2	7	11	36	24	33	100	4	12	35	26	30	100	4	8	35	25	34	100
-2	-2	60	29	6	39	26	4	75	26	5	37	17	4	100	31	4	40	17	4
1	0	36	86	10	21	39	7	38	64	8	21	31	8	34	100	9	21	34	8
6	6	6	11	89	7	13	35	6	12	97	8	13	35	4	9	100	7	14	35
4	-2	31	19	7	98	70	25	34	15	7	95	65	25	40	21	5	100	62	26
14	2	26	34	13	60	85	34	21	27	14	63	88	34	17	34	4	62	100	34
7	2	7	11	36	24	33	100	4	12	35	26	30	100	4	8	35	26	34	100

1 x of 36 Classification Structures

In addition to the entire 36 x 36 matrix, which contains 1,296 values, sections of the matrix are presented below in tabular form as these data relate to the attributes being investigated.

a. The Effect of the Classification Algorithm--WD-2 or WD-3--on the Reliability of the Classification Structure

Since, as was discussed in Paragraph 3b, the algorithms used in the WD-2 and WD-3 programs are different, it was desirable to investigate the degree of similarity between the classification structures that result from their use. Would those different machine procedures yield very different or very similar classification structures? The results, recorded in the last column of Figure 11, list the values of the correlation coefficient as varying from .28 to .63. These figures indicate that there is a slight to moderate degree of similarity between the structures derived by the two classification procedures. These results are in accord with our intuitive expectation, and they reinforce our notion that, even though the inputs are the same, different machine classification algorithms will result in different document groupings--the degree of similarity being dependent upon the similarity of the procedures used. This last statement should perhaps be modified somewhat, for the data seem to suggest that machine classification based upon machine-derived index terms is slightly more reliable than is machine classification based upon concept index terms. However, this is only a tentative formulation,

	A	B	C	D	E	F
	Classification Algorithm Pairings	Type of Indexing	Depth of Indexing	Order	Matrix Pair	Correlation
1	WD-2&3	M	6	01	1-4	.28
2	WD-2&3	M	6	02	7-13	.51
3	WD-2&3	M	6	03	10-16	.53
4	WD-2&3	M	15	01	2-5	.46
5	WD-2&3	M	15	02	8-14	.39
6	WD-2&3	M	15	03	11-17	.63
7	WD-2&3	M	30	01	3-6	.26
8	WD-2&3	M	30	02	9-15	.29
9	WD-2&3	M	30	03	12-18	.34
10	WD-2&3	H	6	01	19-22	.33
11	WD-2&3	H	6	02	25-28	.34
12	WD-2&3	H	6	03	31-34	.40
13	WD-2&3	H	15	01	20-23	.39
14	WD-2&3	H	15	02	26-29	.28
15	WD-2&3	H	15	03	32-35	.34
16	WD-2&3	H	30	01	21-24	.36
17	WD-2&3	H	30	02	27-30	.35
18	WD-2&3	H	30	03	33-36	.35

Figure 11. Correlation of Classification Structures when the Classification Algorithm Is Varied

based upon the fact that the correlation coefficients are slightly higher for machine indexing than for human indexing. These results need to be verified before they can be given much credence.

b. The Effect of Indexing Type on the Reliability of the Classification Structure

This section describes the study of the differences in classification structure caused by human and machine-aided indexing as all other variables are held constant. It will be recalled that human indexing was done by trained librarians from the SDC library staff who were asked to assign appropriate multiple-word concepts as index terms. In contrast, the machine-aided indexing was of the single-word uniterm type.

The table in Figure 12 is arranged to show, in a clear and unmistakable manner, the importance that indexing style--human or machine-aided--has on the structure of the automatically produced classification hierarchy. Column A is the same throughout the 18 rows; the letters M and H simply indicate that in all cases we will be comparing machine and human indexing. The first nine rows of column B indicate that we will initially examine the data generated by classification algorithm WD-2 and then look at WD-3. Column C indicates the depth of indexing and column D the order of input. In column E are listed the pairs of classification structures that meet all preceding conditions (see Figure 9); and in the last column, the values of the appropriate correlation coefficients are recorded.

	A Indexing Type Pairing	B Classification Algorithm	C Depth	D Order	E Matrix Pair	F Correlation
1	M&H	WD-2	6	01	1-19	-.01
2	M&H	WD-2	6	02	7-25	.03
3	M&H	WD-2	6	03	10-31	.07
4	M&H	WD-2	15	01	2-20	-.01
5	M&H	WD-2	15	02	8-26	-.03
6	M&H	WD-2	15	03	11-32	.05
7	M&H	WD-2	30	01	3-21	.01
8	M&H	WD-2	30	02	9-27	.00
9	M&H	WD-2	30	03	12-33	-.01
10	M&H	WD-3	6	01	4-22	.19
11	M&H	WD-3	6	02	13-28	.19
12	M&H	WD-3	6	03	16-34	.16
13	M&H	WD-3	15	01	5-23	.03
14	M&H	WD-3	15	02	14-29	.06
15	M&H	WD-3	15	03	17-35	.14
16	M&H	WD-3	30	01	6-24	.03
17	M&H	WD-3	30	02	15-30	.02
18	M&H	WD-3	30	03	18-36	.02

Figure 12. Correlation of Classification Structures when the Type of Indexing Is Varied

These data are of great significance. They clearly show that, given the same set of documents, machine-aided indexing based upon key words will result in an entirely different distribution of the parent documents in the machine-produced classification structure than would be obtained if the input lists were multiple-word concept terms prepared by skilled humans. Note that we have not said that one structure is better than the other (we discuss utility in Section VI of this report) but only that the structures are significantly different to a degree that most of us would not have anticipated.

All the values in column F are essentially zero, with the exception of the values in rows 10, 11, and 12, which show slight positive correlations. These occur under classifications procedures WD-3 when the depth of indexing is six terms. Under these conditions, there are the greatest similarities--although still slight--between the classification structures derived from human and machine-aided indexing. This finding is consistent with the findings that the effect of depth of indexing is less marked when human index lists of six terms are processed by the WD-3 program.

c. The Effect of Indexing Depth on the Reliability of the Classification Structure

The next question we wish to investigate is whether the number of indexing terms on the list being processed affects the classification structure even though all other variables are counterbalanced.

Another way of phrasing this same question is to ask whether the classification structures derived from 6, 15, and 30 terms would be significantly different from each other.

First let us examine the effect of using indexing lists of 6, 15, and 30 terms that were machine derived, input into the WD-2 program in order #1. Since there are three conditions, taken two at a time, there are three interrelationships--rows 1, 2, and 3 of Figure 13.

These first three rows are interpreted to mean that the hierarchical classification structures derived by using different length lists of terms have an essentially zero correlation, and are therefore not all similar to each other.

But before coming to any overall conclusion, let us examine the next six rows in Figure 13. If the hierarchical arrangement of the documents in a classification structure is <u>primarily</u> dependent upon the length of the index list, then we would expect to find similar results over the three orders of input.

The expected results are borne out by an examination of rows 4, 5, and 6, and rows 7, 8, and 9. Simply varying the size of the list while using machine-aided indexing and the WD-2 program will cause significant changes in the classification structure.

	A	B	C	D	E	F
	Depth of Indexing Pairings	Type of Indexing	Algorithm	Order	Matrix Pair	Correlation
1	06-15	M	WD-2	01	1-2	.18
2	06-30	M	WD-2	01	1-3	.20
3	15-30	M	WD-2	01	2-3	.21
4	06-15	M	WD-2	02	7-8	.30
5	06-30	M	WD-2	02	7-9	.20
6	15-30	M	WD-2	02	8-9	.31
7	06-15	M	WD-2	03	10-11	.23
8	06-30	M	WD-2	03	10-12	.26
9	15-30	M	WD-2	03	11-12	.52
10	06-15	M	WD-3	01	4-5	.24
11	06-30	M	WD-3	01	4-6	.31
12	15-30	M	WD-3	01	5-6	.55
13	06-15	M	WD-3	02	13-14	.31
14	06-30	M	WD-3	02	13-15	.27
15	15-30	M	WD-3	02	14-15	.53
16	06-15	M	WD-3	03	16-17	.29
17	06-30	M	WD-3	03	16-18	.29
18	15-30	M	WD-3	03	17-18	.46

Figure 13. Correlation of Classification Structures when the Number of Machine-Aided Index Terms Is Varied

There is another bit of data that is worth noting: The last line in all three sections--that is rows 3, 6, and 9--has the highest numerical value, which shows the greatest similarity between the classifications based on 15 and 30 terms. This was a very tentative formulation--for certainly the numerical values are not that far apart--but it was worth checking.

We continued to investigate the significance of index lengths to see where these same relationships held when we used the WD-3 classification algorithm. These data are in the bottom half of Figure 13.

The correlation coefficients on lines 10, 11, and 12; 13, 14, and 15; and 16, 17, and 18 are indeed quite similar to the first three subsections of this table, and we concluded that the length of the index list can significantly affect the arrangement of items in an automatically derived classification hierarchy, and that this relationship holds, whatever the order of list processing, or whether the classification algorithm is WD-2 or WD-3.

We have yet to see whether this same phenomenon would hold if the index term lists were derived by skilled librarians rather than machine-aided technique.

Let us examine the data in Figure 14. Note that lines 1, 2, and 3 are exactly comparable to the first three lines in Figure 13, except that Figure 14 is based upon human indexing, while Figure 13 contains machine-aided index lists. Since the values in the first three rows of Figure 14 are slightly lower, it would appear that classification structures based upon human indexing are more subject to variation as the number of index terms per list is increased than are classification structures based on machine-aided index lists.

We checked to see whether this trend continued as we examined additional data, varying only the order of input. The sets of correlation coefficients in the first three sections of Figure 14 are almost identical in their values. This is not surprising, for the only attribute varied was the order in which the lists were presented to the program for processing. At any rate, an examination of these three sets of data lends support to our notion that machine-derived classification structures based upon human-produced index lists are sensitive to the number of terms on the lists--or, stated differently, that different-sized lists will produce dissimilar hierarchical classifications.

There is one other bit of data that merits our attention. In all three of these sections, the highest coefficient was obtained when the classification structures based upon 6-term and 15-term index lists were correlated (lines 1, 4, and 7). This contrasts markedly with the corresponding data in Figure 13, where the highest value was between 15- and 30-term lists.

	A	B	C	D	E	F
	Depth of Indexing Pairings	Type of Indexing	Algorithm	Order	Matrix Pair	Correlation
1	06-15	H	WD-2	01	19-20	.30
2	06-30	H	WD-2	01	19-21	.05
3	15-30	H	WD-2	01	20-21	.11
4	06-15	H	WD-2	02	25-26	.35
5	06-30	H	WD-2	02	25-27	.05
6	15-30	H	WD-2	02	26-27	.12
7	06-15	H	WD-2	03	31-32	.34
8	06-30	H	WD-2	03	31-33	.04
9	15-30	H	WD-2	03	32-33	.09
10	06-15	H	WD-3	01	22-23	.69
11	06-30	H	WD-3	01	22-24	.24
12	15-30	H	WD-3	01	23-24	.33
13	06-15	H	WD-3	02	28-29	.65
14	06-30	H	WD-3	02	28-30	.26
15	15-30	H	WD-3	02	29-30	.30
16	06-15	H	WD-3	03	34-35	.62
17	06-30	H	WD-3	03	34-36	.26
18	15-30	H	WD-3	03	35-36	.34

Figure 14. Correlation of Classification Structures when the Number of Human-Selected Index Terms Is Varied

Turning our attention to the bottom half of Figure 14, we checked to see whether the WD-3 algorithm yielded data that were similar to that obtained by the WD-2 procedures. The lower three sections reveal that they are quite similar to each other and that they contain higher values than those in the upper portion of the figure. Also, the highest values, in the .60's, occur between list lengths at 6 and 15 terms.

The analysis of these tables provided a basis for our answering the question: Do automatically derived hierarchical classification structures based upon index lists containing 6, 15, and 30 terms differ significantly?

The answer was, clearly, "yes"; the size of the index term list affects the classification structure, although the effect is less marked when human indexing to a depth of six terms is processed by the WD-3 programs.

d. The Effect of the Order in Which the Documents Are Input for Processing

The final variable that we wished to investigate was the effect of the input order on the reliability of the classification structure. In the description of the hierarchical grouping program (Paragraph 1), we explained that the program combines documents into groups by searching the similarity matrix for the highest cell value. However, in case more than one cell has the same value, the first cell position is used to form the group. It is thus possible that the order of

input may have an effect on the final clustering structure. To test and evaluate the significance of the input order, three different arrangements were used (see Paragraph 3c). The results are shown in Figures 15 and 16.

The order of input did cause some variation in the final classification structures, but by itself, this was not a very significant factor. It was also clear that this variation was less for the WD-3 than for the WD-2 algorithm. Furthermore, classification structures based upon documents that had been indexed by trained indexers using multiple-word concept terms were less subject to variation than were classification schemes using machine-derived word lists.

One additional point worth noting is that the more terms used to describe the document, the less likely was the classification structure to vary, because differences in document input had no effect whatsoever.

5. A FACTOR ANALYSIS OF THE CORRELATION MATRIX

In collecting data on the reliability and consistency of automatically derived classification structures, we computed a table of intercorrelation for the various classification structures (Figure 10). While individual correlation coefficients were interpreted, and the results discussed in the preceding paragraphs (Paragraphs 4a through 4d), we also analyzed the matrix itself.

	A Order of List Input Pairings	B Type of Indexing	C Algorithms	D Indexing Depth	E Matrix Pair	F Correlation
1	01-02	M	WD-2	06	1-7	84
2	01-03	M	WD-2	06	1-10	78
3	02-03	M	WD-2	06	7-10	90
4	01-02	M	WD-2	15	2-8	58
5	01-03	M	WD-2	15	2-11	53
6	02-03	M	WD-2	15	8-11	48
7	01-02	M	WD-2	30	2-9	74
8	01-03	M	WD-2	30	3-12	57
9	02-03	M	WD-2	30	9-12	61
10	01-02	M	WD-3	06	4-13	86
11	01-03	M	WD-3	06	4-16	96
12	02-03	M	WD-3	06	13-16	87
13	01-02	M	WD-3	15	5-14	84
14	01-03	M	WD-3	15	5-17	85
15	02-03	M	WD-3	15	14-17	89
16	01-02	M	WD-3	30	6-15	87
17	01-03	M	WD-3	30	6-18	91
18	02-03	M	WD-3	30	15-18	95

Figure 15. Correlation of Classification Structures when the Order of Input for Machine-Aided Index Term Lists Is Varied

	A	B	C	D	E	F
	Order of List Input Pairings	Type of Index	Algorithm	Indexing Depth	Matrix Pair	Correlation
1	01-02	H	WD-2	06	19-25	.64
2	01-03	H	WD-2	06	19-31	.60
3	02-03	H	WD-2	06	25-31	.75
4	01-02	H	WD-2	15	20-26	.62
5	01-03	H	WD-2	15	20-32	.86
6	02-03	H	WD-2	15	26-32	.64
7	01-02	H	WD-2	30	21-27	.91
8	01-03	H	WD-2	30	21-33	.89
9	02-03	H	WD-2	30	27-33	.97
10	01-02	H	WD-3	06	22-28	.95
11	01-03	H		06	22-34	.98
12	02-03	H	WD-3	06	28-34	.95
13	01-02	H	WD-2	15	23-29	.83
15	01-03	H	WD-2	15	23-35	.85
15	02-03	H	WD-3	15	29-35	.88
16	01-02	H	WD-3	30	24-30	1.00
17	01-03	H	WD-3	30	24-36	1.00
18	02-03	H	WD-3	30	30-36	1.80

Figure 16. Correlation of Classification Structures when the Order of Input for Human-Selected Index Term Lists Is Varied

The 36 x 36 correlation matrix was factor analyzed using a principal component solution (Harmon, 1967). Ten principal axes were extracted, six of which accounted for 68.6 percent of the total variance. These were rotated orthogonally for simple structure, and the results recorded in Figure 17. Figures 18 through 22 were derived from the rotated factor matrix, but, for ease of interpretation, we show only the significant loadings, arranged in descending order. To the right of the values are listed the attributes of the classification structure or array.

The interpretation of these factors is clear: The attributes that have a significant effect on the similarity of the machine-derived classification structure are primarily the type of indexing (machine derived or human indexing) and the number of index terms used. This analysis is supported by the interpretation of the factors, which have been labeled as follows:

Factor I. Machine Indexing, Long Lists (Figure 18)

Factor IV. Machine Indexing, Short Lists (Figure 19)

Factor III. Human Indexing, Long Lists (Figure 20)

Factor V. Human Indexing, Short Lists, WD-2 (Figure 21)

Factor II. Human Indexing, Short Lists, WD-3 (Figure 22)

Factor VI. Machine Indexing, Long Lists, WD-3 (Figure 23)

Note that instead of having a single factor dealing with human indexing, short lists, we have two--one for each of the two classification algorithms. Note also that the last factor (Machine

	Factor					
Array	I	IV	III	IV	V	VI
1	.296	-.167	.065	.703	.097	-.216
2	.640	-.007	.001	.109	-.067	.002
3	.597	-.080	-.045	.085	.140	.059
4	.013	.228	.002	.780	-.069	.382
5	.658	.068	.061	.056	-.070	.499
6	.255	-.008	.032	.066	-.020	.863
7	.318	-.071	.013	.798	.099	-.207
8	.590	.149	-.018	.250	-.068	.003
9	.702	-.320	-.036	.042	.108	.086
10	.272	-.236	-.015	.846	.065	-.143
11	.784	.019	.020	.119	-.000	.110
12	.627	-.081	-.042	.128	.155	.167
13	.147	.191	-.007	.818	-.045	.275
14	.705	.070	.081	.094	-.062	.479
15	.278	-.010	.016	.056	-.010	.858
16	.050	.186	.008	.813	-.061	.380
17	.697	.115	.066	.087	-.077	.419
18	.271	-.013	.035	.028	.001	.881
19	.061	.254	-.026	-.044	.628	.032
20	.069	.133	.114	.025	.749	-.023
21	-.086	-.089	.832	.048	.188	.107
22	-.035	.869	-.005	.067	.208	.008
23	-.003	.812	.131	.036	.306	-.007
24	.109	.410	.748	-.044	-.144	-.122
25	-.040	.240	-.046	-.014	.701	-.003
26	.016	.081	.144	.021	.687	-.047
27	-.085	-.089	.841	.049	.180	.136
28	-.038	.882	.014	.071	.191	.036
29	-.008	.796	.129	.033	.217	.032
30	.109	.411	.751	-.045	-.139	-.121
31	-.008	.247	-.057	.023	.642	-.020
32	.048	.141	.077	.048	.785	-.016
33	-.083	-.078	.838	.042	.103	.136
34	-.050	.881	.005	.078	.196	.000
35	.036	.780	.174	-.005	.222	.028
36	.109	.411	.751	-.045	-.138	-.121

Figure 17. Rotated Factor Matrix

Array	Value	Attributes
11	.784	WD-2 M 15 3
14	.705	WD-3 M 15 2
9	.702	WD-2 M 30 2
17	.697	WD-3 M 15 3
5	.658	WD-3 M 15 1
2	.640	WD-2 M 15 1
12	.627	WD-2 M 30 3
8	.590	WD-2 M 15 2
3	.579	WD-2 M 30 1

Figure 18. Factor # I: Machine Indexing, Long Lists

Array	Value	Attributes
10	.846	WD-2 M 6 3
13	.818	WD-3 M 6 2
16	.813	WD-3 M 6 3
7	.798	WD-2 M 6 2
4	.780	WD-3 M 6 1
1	.703	WD-2 M 6 1

Figure 19. Factor # IV: Machine Indexing, Short Lists

Array	Value	Attributes
27	.841	WD-2 H 30 2
33	.838	WD-2 H 30 3
21	.832	WD-2 H 30 1
36	.751	WD-3 H 30 3
30	.751	WD-3 H 30 2
24	.748	WD-3 H 30 1

Figure 20. Factor # III: Human Indexing, Long Lists

Array	Value	Attributes
32	.785	WD-2 H 15 3
20	.749	WD-2 H 15 1
25	.701	WD-2 H 6 2
26	.687	WD-2 H 15 2
31	.642	WD-2 H 6 3
19	.628	WD-2 H 6 1

Figure 21. Factor # V: Human Indexing, Short Lists, WD-2

Array	Value	Attributes
28	.882	WD-3 H 6 2
34	.881	WD-3 H 6 3
22	.869	WD-3 H 6 1
23	.812	WD-3 H 15 1
29	.796	WD-3 H 15 2
35	.780	WD-3 H 15 3

Figure 22. Factor # II: Human Indexing, Short Lists, WD-3

Array	Value	Attributes
18	.881	WD-3 M 30 3
6	.863	WD-3 M 30 1
15	.858	WD-3 M 30 2
5	.499	WD-3 M 15 1
14	.479	WD-3 M 15 2
17	.419	WD-3 M 15 3
4	.382	WD-3 M 6 1
16	.380	WD-3 M 6 3
13	.275	WD-3 M 6 2

Figure 23. Factor # VI: Machine Indexing, Long Lists, WD-3

Indexing, Long Lists, WD-3) is partially redundant with the first factor and begins to divide that first factor in accordance with the classification algorithm used. These findings are consistent with the statement made earlier that the variables that contribute most to the structure of the machine-derived classification system are type of indexing and the number of index terms.

6. SUMMARY OF RESULTS AND CONCLUSIONS

The experiments described in Paragraph 4 were designed to investigate the reliability and consistency of the ALCAPP programs for automatically deriving classification hierarchies as four selected attributes were varied under controlled conditions. These attributes were:

(1) The classification algorithm.

(2) The type of document indexing.

(3) The depth of indexing.

(4) The order of document input for processing.

A total of 36 different combinations were studied, and 36 classification structures derived. In order to investigate the consistency or similarity of these structures, each classification was compared with every other one and the results of these comparisons summarized in a matrix of intercorrelations (Figure 10). This matrix provides the data for analysis and interpretation.

a. The Importance of the Classification Algorithm

Two classification algorithms were used, and these are designated WD-2 and WD-3. The WD-2 procedure results in a classification structure that is relatively symmetric and consists of a few main clusters (Figure 4). The WD-3 algorithm creates some main clusters of similar documents plus small clusters of a few documents each, and finally, some clusters of two, three, or even single documents. The result is an asymmetric hierarchy (Figure 6).

Clearly, the classification structures are going to be somewhat different, but the question being investigated was whether the distributions of the documents within the two hierarchical classification structures were similar--that is, would documents that were put in the same cluster by one algorithm also tend to be close together in the classification structure created by use of the other algorithm?

The results of the comparisons showed that there was a slight to moderate degree of similarity between the classification structures that were derived from the WD-2 and WD-3 algorithms, and this is what we would expect.

The data from the factor analysis support this conclusion, and amplify it, by indicating that the classification algorithm has a greater effect when the input lists consist of relatively few human-derived concept terms. Under these circumstances, the WD-2 algorithm would tend to force the document into a cluster, while the WD-3 algorithm would tend

to keep that document list separate and distinct--thus creating a greater degree of dissimilarity than would be obtained under other combinations of attributes.

At any rate, the mathematical and logical techniques used for making a classification structure had a moderate effect on the similarity of the resulting structures.

b. The Importance of the Type of Document Indexing

Two types of indexing were used in constructing surrogate lists for input to the classification programs; these were:

(1) Concept indexing done by trained librarians.

(2) Key-word indexing using machine-aided techniques.

This experiment provided clear evidence of the fact that these different indexing techniques would result in machine-produced classification structures that had little resemblance to one another. This is a most significant finding, for it states that regardless of the other factors involved, document subject groupings differ, depending upon whether the index terms used are uniterms or pre-coordinated subject headings.

c. The Importance of Depth of Indexing

Each document was indexed by 6, 15, and 30 terms. The question being investigated was whether, other things being equal, classification

structures would differ significantly as a result of the number of index terms used.

The results of the individual comparisons and of the factor analysis clearly indicated that classification structures derived by using long lists of index terms differ significantly from the structures derived by using short lists.

Again, this is a rather important finding, for it points to the danger of intermingling depth indexing with shallow indexing when organizing document collections.

d. The Importance of the Order of Document Input

The particular procedures used in creating clusters of documents can be affected by which documents are used for creating the original groupings. To determine the importance of this variable on the similarity of the resultant classification structure, the input order was varied and the effects studied.

The results showed that the effect of input order was not very significant, and that the classification structures derived by using different orders of input were quite similar.

From a practical point of view, this finding is important because, if the processing order can be ignored (as indeed it can), the classification algorithm can be simplified.

SECTION V

MEASURING THE STABILITY OF AUTOMATICALLY-DERIVED DOCUMENT GROUPINGS

A manual classification system relies on human ingenuity to insure flexibility as new and related items are added to the document collection and to do this in such a way that the stability of the original structure is maintained. Nevertheless, all classification systems tend to become rigid over a period of time, and when significant changes are made in the character of the collection, their efficiency is reduced, for the systems cannot be revised radically except at great cost.

One of the unique advantages of automated document classification is that the entire collection of materials can be reclassified periodically and relatively inexpensively. However, it is important that even with reclassification a certain stability and consistency be maintained. Documents that have been previously grouped together should not, in the reorganized system, appear to be unrelated.

Reclassification by means of the ALCAPP cluster-finding programs has been demonstrated to be relatively simple and inexpensive. Cost goes up linearly with the number of documents being processed rather than exponentially, as is the case of some procedures. However, the stability of the classifications needs to be evaluated. In order to do so, two series of experiments were designed. The first set investigated the sensitivity of the ALCAPP clustering algorithm to changes in initial

cluster assignments, and the second investigated the stability of the classification system as the size of document collection is increased incrementally.

1. DESCRIPTION OF THE CLUSTER-FINDING ALGORITHM

The data on which the program operates are a set of word lists derived from documents. These words may be assigned by human indexers or by computer. In these experiments, the basic document collection consists of 600 abstracts, and each of these is reduced to a list of 30 terms by machine methods, as described in Section III, paragraph 2.

The ALCAPP cluster-finding algorithm is an iterative procedure, which starts with an arbitrary assignment of documents to an arbitrary number of clusters. Then with each iteration, documents judged to be similar on the basis of the word lists are grouped together, and previously unassigned documents are added to the clusters. A detailed description of our procedure follows.

Input Stage: We began by choosing a reasonable number of clusters into which the total collection can be divided. The actual number of categories in the final classification scheme may be less than this upper bound, depending on the differences in content among the documents. In all of these experiments the initial number of categories was set at ten.

After the number of categories to be used for the initial iteration was determined, a set of documents was assigned to each cluster, but no document was assigned to more than one cluster. Twenty documents were so assigned. Both the number of documents and their selection were arbitrary. We wished to choose a reasonable-sized sample, but at the same time we wished to maintain a large pool of unassigned documents, for these help to differentiate and separate the categories. As will be seen, the program shifts documents from their originally assigned categories and brings in new documents from the unassigned pool.

First Iteration:

The inidividual word lists in each cluster were combined into a single composite list or dictionary of all of the terms and their frequency of occurrence. The dictionary list was then rearranged so that the most frequently occurring words were listed on top and the other words followed on a descending order of frequency.

Weights were assigned to each term on the list. The highest weight assigned was equal to the number of documents, or individual word lists, that make up the composite dictionary for that cluster. In this case twenty word lists were used (N=20); therefore, the most frequent word in the dictionary was assigned a weight of twenty. The next most frequently occurring word was assigned a weight of N-1 or 20-1 = 19, and so on until each word had been assigned a weight.

Four other constraints were imposed on the program for assigning weights:

(a) The highest value, or greatest possible weight that can be assigned to any term was 63. A cluster set that contains more than 63 document lists will nonetheless have 63 as the maximum weight.

(b) All terms with the same frequency were assigned the same weight.

(c) No term was assigned a negative or zero weight. Thus, if more frequency classes exist than the highest weight, all lower frequencies will be assigned a weight of 1.

(d) Words that occurred only once and thus had a word frequency of one were automatically assigned a weight of 1.

An abbreviated example of weight assignment is shown in Figure 24.

The first iteration concluded with the assignment of weights to each term in the dictionary list. The resulting list of weighted terms was called the cluster profile; one profile was constructed for each cluster.

Second Iteration:

At the start of the second iteration, a cluster profile existed for each cluster or category. This profile consisted of a list of all the terms appearing in the document word lists in a given cluster, plus their assigned weights. Each cluster had its own profile.

LIST #: 1	2	3	4	5
information document technical requirement science	information retrieval result published analyzed	computer system retrieval information index	education information need description retrieval	research technical information science system

Five Document Term Lists in One Cluster

TERM	FREQUENCY	WEIGHT
information	5	5
retrieval	3	4
system	2	3
science	2	3
technical	2	3
requirement	1	1
result	1	1
published	1	1
analyzed	1	1
index	1	1
computer	1	1
need	1	1
education	1	1
description	1	1
research	1	1

Composite Dictionary and Cluster Profile

Figure 24. Terms Lists and Cluster Profile Weight Assignments

Computer processing for the second iteration began by assigning a score on each profile to every document in the entire collection. If there were ten categories, each document was assigned ten profile scores. A profile score is the arithmetic sum of the weights assigned to terms in a profile that occur in the documents' list of index terms. A ratio between the highest profile score and the next highest score was also computed. The document was then tentatively assigned to the cluster profile on which it received the highest score. It is at this point that initially created categories can disappear. This, indeed, happened in the experiments described below. Of the 600 documents in the collection, no document received its highest score in a particular category. As a result, no documents were assigned to that category; so instead of ten profile clusters there were only nine.

A list was made of the documents assigned to each category. This list was sorted on the profile score ratio that had been computed previously, and the document identification numbers rearranged, so that the one with the highest ratio value appeared on top and the rest were listed in descending order. The top N + 1/2N documents were assigned to each cluster and all other documents were listed in the unassigned pool. N is the number of documents assigned to a cluster in the previous iteration.

By limiting the number of new documents that could be assigned to a cluster in any one iteration to 1/2 N, we could separate the clusters into distinct subject categories and add new documents gradually.

It is perhaps obvious that although no more than N + 1/2 N documents can be assigned to a category at each iteration, this does not mean that many documents will be assigned. Each document receives many profile scores and is tentatively assigned to the category in which it has the highest score. Clearly, more documents could be assigned to one category than to another.

Subsequent Iterations:

The iterative process was repeated. New document profile scores were computed for all the documents in the collection, together with their appropriate ratios. Tentative assignments of documents was made to the most likely category; these were re-sorted by ratio score, additional documents added to the category, etc.

The iterative process continued until (a) every document had been assigned to a category, and (b) the new set of clusters was exactly the same as the set obtained from the previous iteration. That is to say, the clusters were stable, the iterative process converged, and no document changed cluster assignment from one iteration to the next.

In the clustering experiments described below, the algorithm was modified slightly to provide an additional constraint, in order to prevent one cluster from being assigned all the documents in the set. This modification was necessary because of the essential homogeneity of the collection, e.g., all the documents were on the subject of information science. The algorithm was modified so that no cluster could be assigned more than 90 documents, until there were no changes in cluster assignment from one iteration to the next, but before all documents in the collection had been assigned to a cluster.

2. DETERMINATION OF THE SENSITIVITY OF THE ALCAPP CLUSTERING ALGORITHM TO CHANGES IN INITIAL CLUSTER ASSIGNMENTS

In devising classification schemes, be they manual or automatic, one finds that the character of the original set of documents plays a disproportionately important role in determining the nature of the subject categories into which the rest of the documents in the collection will be divided. This is perhaps especially true when using the ALCAPP algorithm, which begins with the arbitrary assignment of a number of documents to each cluster. Yet, ideally, it would be desirable that the final classification be the same regardless of which documents were used in the original cluster assignments. These experiments were designed to determine how far reality departs from this ideal.

a. Purpose

The purpose of this experiment was to determine the sensitivy of the final classification structure to differences in the initial assignment of documents to clusters.

b. Method

The experimental data set consisted of 600 documents and their surrogates of 600 word lists, each containing 30 key-word terms derived by machine analysis of the document abstract. At the start of the ALCAPP processing procedures, 10 categories were created and 20 documents assigned to each category. Then the program divided the collection into clusters, as described in the previous section. The documents in the initial cluster were assigned randomly, the only constraint being that a document could be assigned to a starting cluster only once in the entire experiment.

The clustering procedure was repeated three times, creating classification structures A, B, and C. In all cases, one category was eliminated by the program; thus, all three structures contained nine categories each. Since there was no reason to expect that any given cluster in one classification would correspond to any particular cluster in a second classification, each cluster in a classification was compared to every cluster in the second classification. The comparison consisted simply in noting the number of documents that the clusters from differing classifications had in common. Hence, we arrived at a 9 x 9 matrix for the comparison of any classification to any other. Since the number of documents in any cluster was known, it was possible to compute the expected value of any cell in the matrix,

assuming only random similarity of document assignments. By comparing the expected values to the observed values, both chi-square and phi could be computed for the entire matrix.

The three matrices are shown in Figures 25, 26, and 27.

c. Findings and Interpretations

For a matrix of this size, the number of degrees of freedom is 64 and the expected value of χ^2 is 128. Clearly, the observed values of 908, 946, and 1263 are significant beyond any chance expectation.

Phi is an index roughly equivalent to a correlation coefficient, with minimum zero and an unpredictable maximum in the neighborhood of 1. The average value observed here, .464, confirms what a visual inspection of the similarity matrices suggests: that while generally no one cluster in a classification can be unambiguously assigned to a given cluster in a second classification, the documents in the first cluster are not distributed randomly among the clusters in the second classification; the bulk of the distribution tends to be concentrated in two or three clusters.

The three clasaifications structures are only moderately similar, but the similarity that exists is not the result of chance; it is statistically very significant.

Categories in Classification A	Categories in Classification B: 1	2	3	4	5	6	7	8	9	Total
1	43	4	1	0	5	2	1	0	5	61
2	3	20	6	4	3	17	0	1	2	56
3	17	5	19	9	3	10	9	20	8	100
4	0	3	5	11	0	11	0	3	7	40
5	5	6	7	3	10	17	1	5	15	69
6	8	5	2	3	10	12	10	3	17	70
7	2	4	1	0	0	35	0	0	6	48
8	18	15	7	5	5	6	15	5	24	100
9	0	4	3	12	0	3	1	33	0	56
Total	96	66	51	47	36	113	37	70	84	600

$\varphi' = .435$

$\chi^2 = 908.3$

Figure 25. Matrix Comparing the Categories in Classification A with B

Categories in Classification A

Categories in Classification C	1	2	3	4	5	6	7	8	9	Total
1	22	6	12	1	4	11	2	17	2	77
2	10	6	20	0	10	9	1	10	0	66
3	20	2	28	2	6	22	5	39	4	128
4	3	2	2	2	15	5	10	0	1	40
5	2	5	3	6	6	5	27	14	0	68
6	2	9	9	8	13	2	0	8	9	60
7	0	13	3	12	5	2	3	3	2	43
8	0	3	1	6	4	3	0	3	6	26
9	2	10	22	3	6	11	0	6	32	92
Total	61	56	100	40	69	70	48	100	56	600

$\rho' = .444$

$\chi^2 = 946.3$

Figure 26. Matrix Comparing the Categories in Classification A with C

Categories in Classification C	Categories in Classification B: 1	2	3	4	5	6	7	8	9	Total
1	21	5	6	2	9	7	1	6	20	77
2	19	6	13	1	7	5	4	3	8	66
3	47	5	4	2	7	10	23	7	23	128
4	4	4	1	0	7	19	1	1	3	40
5	1	19	2	4	2	28	2	2	8	68
6	2	8	13	8	3	9	0	8	9	60
7	1	6	1	4	0	22	1	3	5	43
8	0	2	0	11	1	1	2	5	4	26
9	1	11	11	15	0	12	3	35	4	92
Total	96	66	51	47	36	113	37	70	84	600

$\varphi' = .513$

$x^2 = 1263.5$

Figure 27. Matrix Comparing the Categories in Classification B with C

Nevertheless, based upon the obtained results, it is recommended that if the ALCAPP automatic classification programs are to be used in a practical way, then some sort of "seeding" process must be used for the initial assignment of documents to clusters. The probable success of such an approach is suggested by a second experiment described below.

3. DETERMINATION OF THE SENSITIVITY OF THE ALCAPP CLUSTERING ALGORITHM BY THE ADDITION OF DOCUMENTS TO A PREVIOUSLY CLASSIFIED SET

All classification systems are organized on the basis of an initial collection of documents. Afterwards, even though many new documents are added to the collection, the original set of categories is expected to be stable. In manual systems, logical organizing principles are used and stability is assured by the ingenuity of human classifiers. In an automated system, the basis of classification is the similarity of the words used in the document or assigned to the word list characterizing that document. Instead of being able to rely on the ingenuity of a trained librarian, we must rely on the logic of the computer program. Granted that there are differences in procedures as well as advantages and disadvantages on both sides, the fact remains that any document classification system must be reasonably stable over time and as new documents are added to the collection. If there is no stability, it would be impossible to learn to use the system, and the advantages of classification would be nullified.

a. Purpose

The purpose of this experiment was to determine just how sensitive the ALCAPP classification algorithm was to the addition of new documents. A classification system is stable when additions can be made to the already established classification categories, and the documents that have been previously assigned to one cluster will not be reassigned to a different cluster.

b. Method

Using the cluster-finding algorithm, 500 of the 600 documents in the collection were classified into 8 clusters. Note that in this experiment, in contrast with the one discussed in paragraph 2, only eight clusters remained, rather than nine, although both experiments started with an initial assignment of ten clusters. The probable reason for this difference is that in the present instance, 500, not 600 documents were classified.

The distribution of the 500 documents in the 8 clusters constituted the initial cluster assignments. To test the stability of this classification, 10, 25, 75, and 100 documents were added to the original 500 and the program was iterated until the standard termination conditions were reached--that is until all the documents had been assigned, and no document changed assignment from one cycle to the next. We compared each of the resulting classifications to the 'baseline' classification by counting the number of the original 500 documents that had been reassigned to a different cluster. In addition, each classification was

Compared to every other classification in order to compute stability as a function of the number of documents added.

c. Findings and Interpretations

The first results of the experiment are described in two tables. Figure 28 records the number of the original 500 documents that changed assignments when 10, 25, or more documents were added to the collection. Figure 29 is in the form of a diagonal matrix and records, for each pair of conditions, the proportion of documents that had different assignments in the two classifications considering only those documents that the pair had in common.

A cursory glance at the data in both figures reveals that a significant difference occurred when 75 documents were added. The number of changing assignments jumped from 18, when 50 new documents were added, to 235, when 75 documents were added, and the percentage change went from about 4 percent to approximately 45 percent.

Two possibilities could account for this dramatic change: either there is a major difference in the content of the last 25 documents added, or the algorithm itself becomes less stable when more than 10 percent of the original data base is added at one time. Since the documents were selected at random from a homogeneous data base, it is unlikely that there is a real difference in the content of the documents. This being the case, the workings of the algorithm itself needs to be studied and evaluated.

NUMBER OF DOCUMENTS BEING CLASSIFIED	NUMBER OF DOCUMENTS THAT CHANGE CLUSTER ASSIGNMENT
500	0
510	20
525	31
550	18
575	235
600	172

Figure 28. Number of Documents Changing Cluster Assignments

	500	510	525	550	575	600
600	.34	.33	.31	.34	.38	.00
575	.47	.47	.45	.47	.00	
550	.04	.03	.05	.00		
525	.07	.07	.00			
510	.04	.00				
500	.00					

Figure 29. Percentage of Documents that Have Changed Cl ıster Assignments

An additional experiment was designed to compare the stability of the classification system when 75 new documents--more than 10 percent--were added to the 500, as compared with the addition of the last 25 documents to a starting classification containing 550 documents. The first set of conditions was simply a restatement of the previously followed procedure in which 235 documents changed cluster assignments. In the second technique, the starting classification contains 550 documents and not 500. To these 550 documents, 25 were added--less than 10 percent--and the entire group of 575 was reclassified. These two procedures, each containing the same number of documents in the final classification, could then be compared.

The results of this experiment are contained in the two matrices illustrated in Figures 30 and 31. The question being investigated was: How many of the original 500 documents change cluster assignments when 75 new documents are added to an initial classification structure containing 500 documents, as compared with adding the same 75 documents in two stages, first 50 and then 25? (Fifty was chosen as the starting point since the set of 550 resulted in the last stable configuration.)

Note that the number of documents changing category assignments in Figure 29 is large, while in Figure 30 relatively few documents change cluster assignments. This is interpreted to mean that, using the ALCAPP algorithms, no more than ten percent of the new documents should be added to an existing classification structure at any one time. Under these conditions, the basic classification structure remains reasonably stable.

Cluster Distribution of the Experimental Set of 500 Documents: Total Set = 500

New Cluster Distribution when 75 ther Documents Are Added	1	2	3	4	5	6	7	8	Total
1	1	2	3	4	5	6	7	8	Total
2	76	2	1	21	0	6	13	17	136
3	4	33	0	30	0	0	20	1	88
4	0	0	11	0	0	0	0	0	11
5	1	6	1	21	0	1	7	3	40
6	0	0	1	0	26	1	8	3	39
7	5	0	0	19	0	47	11	6	88
8	0	1	1	6	4	0	14	0	26
	7	0	2	3	0	2	21	37	72
Total	93	42	17	100	30	57	94	67	500

Figure 30. Matrix Showing the Number of Documents Changing Categories when the Number of Additional Documents Exceeds Ten Percent

New Cluster Distribution when 25 Other Documents Are Added	Cluster Distribution of the Experimental Set of 500 Documents: Total Set = 550								
	1	2	3	4	5	6	7	8	Total
	90	0	0	1	0	0	3	1	95
	0	41	0	0	0	0	0	0	41
	0	0	16	0	0	0	0	0	16
	3	1	0	95	0	1	3	0	103
	0	0	0	0	28	0	3	0	31
	0	0	0	1	1	56	9	1	68
	0	0	1	3	1	0	74	1	80
	0	0	0	0	0	0	2	64	66
Total	93	42	17	100	30	57	94	67	500

Figure 31. Matrix Showing the Number of Documents Changing Categories when the Number of Additional Documents Is Less than Ten Percent

4. DESCRIPTION OF THE CLUSTERS

As an adjunct to these experiments on stability, one of the members of the project staff examined the initial classification of the 500 documents to see if the clusters seemed "reasonable," i.e., whether there was a unifying theme shared by the documents classified in the same category. It was recognized at the outset that this process is highly subjective, and it was undertaken only to make some estimate of the reasonableness of the classification.

The results of this perusal are both satisfying and disappointing; the categories make sense but they are not cohesive. Most of the clusters contained a "core" of documents that were indeed highly related and could sensibly be classified together. On the other hand, two effects were noted that are not reasonable. First, many of the documents in a cluster, say 10 to 20 percent, seem misplaced in the sense that they would appear to fit more appropriately in another of the 8 clusters. Second, certain topics that seem as though they ought to form distinct clusters do not, and are scattered through all the clusters. Examples of this latter case are:

(a) documents related to "artificial intelligence";

(b) documents related to "writing style";

(c) documents related to "machine translation."

One possible explanation for this effect is that there are relatively few documents in these categories--20 to 25. Nonetheless, one could

hope that all documents pertaining to a given topic might have been assigned to the same cluster.

A description of the eight clusters follows, but in these interpretations only the "core" subject area or areas are described, with some indication as to their purity:

Cluster 1: 42 documents:

Automated, computer-oriented information retrieval systems. Fairly cohesive cluster. Oddly enough, a fair number of documents pertaining to medical information retrieval systems, which might have fitted better in Cluster 2, wound up here. The choice is fairly arbitrary, in that the medical systems described are machine-oriented.

Cluster 2: 94 documents:

Descriptions, or descriptions and evaluations, of working IR systems. ("Current awareness," "Documentation Dissemination," etc.) Methods of evaluation of IR systems.

Cluster 3: 57 documents:

Library automation--shelf lists, document control, accessions, etc.
Library cataloging operations--manual or machine.
Impurities: "Automatic text processing";
Documents relating to "costs."

Cluster 4: 67 documents:

Technical communication;

Communication networks;

A reasonably homogeneous cluster.

Cluster 5: 100 documents:

A fairly mixed group containing documents on reproduction methods, publication, hardware descriptions, and chemical IR systems.

Cluster 6: 93 documents:

Educational libraries (i.e., various school libraries);

The education of librarians (library school curricula, etc.);

Professional aspects of librarians;

Specialized Information Centers (medical, agricultural, etc.).

Cluster 7: 30 documents:

Document representation--thesauri, indexing classification, etc.;

A fairly cohesive group of documents.

Cluster 8: 17 documents:

No easily discernible pattern, but generally concerned with representation methods--abstracting and indexing.

5. SUMMARY OF RESULTS AND CONCLUSIONS

A series of experiments was conducted to measure the stability of automatically derived document groupings. The first experiment was designed to determine the sensitivity of the clustering algorithm to changes in the initial assignment of documents made at the start of the program. Three classification structures were derived using different starting assignments. These were compared and found to be only moderately similar, which indicates that the algorithm is sensitive to changes in initial clustering assignments. It is recommended that if the ALCAPP automatic classification programs are to be used in a practical situation, the documents selected for the initial cluster assignments be selected with a view toward achieving a reasonable cluster separation.

The second experiment was aimed at determining the stability of the classification structure as new documents were added to the collection. When a larger percentage of documents were added, the algorithm was not stable. It is therefore recommended that in a practical situation no more than ten percent of new documents be added to the existing classification structure at any one time.

Finally, the documents in the clusters were examined to determine whether the clusters appeared to be cohesive and reasonable from a content analysis point of view. The results show that, while the automatically created clusters are statistically reliable and definitely not random,

the grouping of documents by content is imperfect. If the automated classification structure is to be used for manual search, as well as in a computer retrieval system, we recommend that the clustering algorithm be used to provide the initial rough grouping of the documents that can then be fairly easily modified and made more rational by a trained librarian.

SECTION VI

MEASURING THE UTILITY OF AUTOMATED DOCUMENT CLASSIFICATION HIERARCHIES

A classification system is designed to help people locate documents that are relevant to their interests, and to do so efficiently. If no classification system is available, one has to make a serial search through the entire collection looking for a given subject, a given title, or a given author. Such a comprehensive search is time-consuming, and the usefulness of the classification system is shown in its improvement in the speed and/or accuracy of the retrieval process.

The fact that a classification system can reduce the time required to search a data base is inherent in the logic built into the search strategy. By dividing the total document collection into sections, only those categories relevant to the request are searched and all other portions of the data base are ignored; thus, search time is reduced.

While it is obvious that one can lessen search time by searching fewer documents, one may not be searching only the relevant portions of the data base: the classification system thus serves no useful purpose. On the small data base being used for these experiments it was impossible to make a significant saving in search time, or even to demonstrate how the automatic classification programs divided the collection into clusters that make search and retrieval more efficient. The value of the clustering technique must be tested and demonstrated indirectly. This can be done by comparing people's judgments of the content of the original

document when only the index terms that characterize that document are known and when both the index terms and the classification category to which the document belongs are known. If it can be demonstrated that the accuracy of judging document content improves when documents are classified as well as indexed, then one can infer that classification is an aid in judging the relevancy of a document surrogate and, thus, an aid in document retrieval.

1. COMPARING DOCUMENT REPRESENTATIONS[1]

There is a need for better answers to one question of long-standing interest to persons trying to improve document searching systems: Given a proposed revision in a document-representation technique, how can it be determined whether the proposed change will effect an improvement? A very important part of the answer to this question depends on whether the proposed revisions will actually result in more adequate representations of the documents and the information requirements statements input to the system. Thus, the empirical methods used to test the adequacy of representations are important in guiding the evolution of document-searching methods, and this means that such testing methods ought to be scrutinized regarding their strengths and weaknesses and their potentialities for yielding additional insights into the processes of document representations and searching.

---

[1]The investigator wishes to acknowledge the contribution of Richard Weis to these utility experiments and, particularly, to the discussion that follows on document representations, modeling and scaling, as well as his help in the statistical analysis of the results.

Basic to the idea of adequacy of representation is the notion of the representation process itself. The concept of representation is deceptive in its apparent simplicity, for there is no widely accepted consensus as to the precisely defined limits of this process. One issue is the purpose of the representation. The purpose may be, for example, to inform the user about the contents of a document; to indicate what the contents are; to provide the basis of an accurate, sensitive search for stored documents; to allow the user of a representation to make the same interpretations that he would make if given the full document; or to allow the user of representations to make the same distinctions that he would make between the full documents.

It is important to note that these purposes are not the same and that representations made with one of these purposes in mind may not necessarily fit a different purpose. For this study, we chose to look at representations in terms of their ability to allow the user to make the same distinctions between representations that he would make between the full documents.

The basic purpose of the utility study was to evaluate how useful a selected set of document representations would be as an aid in the retrieval function. The representations used were: a set of index terms produced by the computer, and the computer-produced index terms coupled with each of two types of classification produced by the Hierarchical Clustering Program. The details of the production of these

representations are discussed elsewhere in this report (Section III). The analytic techniques used in part of the study are described in Paragraph 3, Data Analysis. However, the analysis and the conceptual model are so interrelated that it is necessary to discuss both. We shall first sketch out a psychological model.

a. Psychological Model of Multiattributed Objects

In the discussion to follow, a stimulus object refers to a thing in real world; its attributes are the things that describe the object. A stimulus, on the other hand, refers to a construct, roughly the set of values of the attributes of a stimulus object.

Every stimulus object has an uncountable number of attributes; however, in making distinctions between objects, only a relatively few are involved. These are determined by the setting or context in which the comparisons are being made. Clearly, they can include only attributes on which the objects differ. Others, although the objects differ on them, will have no relevance to the comparisons. Further, some attributes may be more important than others; that is, they may loom larger in the comparison. For example, in the document area, grade of paper may be irrelevant and writing style, though relevant, may be secondary to other considerations.

It is an assumption of the model that the objects are measurable with respect to their attributes; in other words, it is possible to make numerical assignments to the objects that reflect the 'amount' of the attribute they have.

Since the objects may vary independently on each of their attributes, a spatial model with the attributes as orthogonal axes, forming a basis, is a natural extension.

In such a model, the stimuli are points in the space and the projections of the points on the basis vectors are the values of the stimuli on the related attributes.

The similarity of stimuli is, then, a function of the distance between them. The form of the distance function is not specified completely, and any distance function that satisfies the Minkowsky inequality is a possible candidate. Most of the early work with this model assumed a Euclidean distance function. In this work, the Euclidean model was used as a first approximation. For a detailed discussion of the forms of the metric and the psychological interpretation of metrics other than Euclidean see Shepard's article in the Journal of Mathematical Psychology (1965).

b. Multidimensional Scaling

If one has all the stimulus objects' scale values on all the relevant dimensions, it is a simple computational task to determine the distance between the points. However, except in a few perceptual domains that have been extensively studied, for example, color vision (Helm, 1959, and Helm and Tucker, 1962), this information is not generally available. Yet it is possible for judges to scale objects in terms of their similarity without reference to particular attributes. It should then be possible to recover the underlying dimensions.

There are two approaches to the analysis of similarities; the older derives from factor analysis, the newer from regression analysis. The factor analysis approach depends on the ability to obtain from a set of similarity judgments a matrix that can be factored.

The newer methods are all related to an algorithm devised by Shepard (1962a, 1962b). The rationale for the Shepard procedure comes from an interesting proof, by Abelson and Tukey, that nonmetric aspects of the data can very closely determine a metric function.

Abelson and Tukey (1959, 1963) show that, if a set of data can be fitted to a so-called ordered metric, that is, roughly, a ranked sequence in which also the first differences are ranked, there is a strategy that will fit a metric to the data that will correlate very highly with the 'true' metric. To paraphrase their finding, the constraints implied by the ranking of the first differences are such that the possible positions of a given point that will preserve the ordering are all very close, so that only a very limited class of distance functions can satisfy the inequalities implied by the ordered metric and that any two of these distance functions will be very highly correlated.

The Shepard method and related methods assume that the similarity judgments are at least monotonically related to the distance function. Consider a regression problem where one variable is the rank of the distance between each pair of points and the other is the

distance; the problem is to find a distance function that minimizes the mean square error in the distance and preserves the order of the distances.

The Shepard algorithm starts with an arbitrary configuration of the points in a multidimensional space. Starting from one point it 'looks at' every other point and decides if the distance between them is too small or too large. It attaches a vector to each point to correct the discrepancy and finally takes the resultant of all the vectors attached to each point as the direction in which to move the point. It takes, as the distance to move the points, a fraction of the length of the resultant, so that the configuration will approach a fit slowly. It then moves all the points; such a move is called a jiggle.

This is the spirit of the three best known methods of multidimensional scaling--Shepard (op. cit.), Kruskal (1964a, 1964b), and Torgerson and Meuser (1962). They differ in detail, especially as to the method of reducing the dimensionality of the configuration. The Torgerson program, which we used, performs the jiggling in successively lower dimensional spaces. The lowest dimensional configuration that meets the goodness-of-fit criterion of the investigator is the one used. Torgerson gives a guideline for goodness of fit but the ultimate criterion is replicability. The Torgerson program also performs the factor analytic procedure at least once to obtain the initial configuration.

The multidimensional scaling model essentially refers to a single individual's perceptual space. To improve the stability and reliability of such an analysis it is common practice to combine the judgments of several individuals into a consensus judgment. This procedure is not without risk. To be valid all the individual's conceptual spaces must be very similar. For example, in the document case, one person may make all his judgments of similarity on the basis of contents of the document while another individual may make his judgments on the basis of writing style. To combine such judgments would possibly violate the assumptions of the model.

Therefore, an analysis of the judgments is performed, a so-called points-of-view analysis (Tucker and Messick, 1963). This analysis is essentially a Q-type factor analysis performed on the cross products of the judgments rather than on correlations. The factors isolated by this procedure roughly correspond to possible bases for making the judgments. A single individual may make his judgments from some combination of bases or points of view. Therefore, each judge receives a 'score' on each factor that indicates how much of that point of view entered into his judgments. Judgment matrices for various 'ideal or hypothetical' judges can be formed by taking linear combinations of the original judgments, using as weights specific combinations of scores on the various factors. In this fashion, it is possible to construct judgment matrices for hypothetical judges

that are not represented in the sample. In our work we did not make much use of this facility, simply because we were not particularly interested in "ideal" judgments. We were interested in finding the judgment matrix that best represented our sample of judges--what we might call a consensus judgment. Our concern was, not to find a large number of points of view, but to assure ourselves that our data were not contaminated by improperly combining variant points of view. To this end, we inspected the cross plots of persons' factor scores, looking for clusters that we had to treat as different points of view. Also we had to judge, from the size of the largest root of the cross product matrix, if we could safely assume that a single point of view would accurately represent our sample of judges.

2. PURPOSE AND METHODOLOGY OF THE UTILITY STUDY

In the utility study, the purpose was to evaluate the effects of automatic index term assignment and automatic classification on document representation. Recall that we defined the 'goodness' of a document representation (surrogate) in terms of how well the representations allow a user to make the same judgments about the documents that he would make given the full text. The judging procedure is an arduous task and experience dictates that subjects can perform the task on at most 12 or 13 full-text documents (this takes about three hours); for representations, the judging process goes considerably faster, taking about one to one and a half hours.

Three sets of documents were selected from the 52 documents used in the previous studies. The first of these was selected so that the documents were rather uniformly distributed among the clusters derived by the WD-2 method; the second set was similarly constructed, using the WD-3 clusters; the third set was chosen on the basis that they had all been assigned to the same category in Documentation Abstracts and yet fell into different clusters on both the WD-2 and WD-3 classifications. It was possible to select these three sets so that each had a subset of six documents in common, thus providing a common core for comparison.

The subjects were divided into four groups, and each group performed a judgment task on each set of documents or corresponding surrogates. For ease in labeling, a code was used to identify the type of representation as follows:

A = the full text of the document

B = lists of machine-derived index terms only

C = lists of human-prepared index terms only

D = lists of machine-derived index terms plus WD-2 classification structure

E = lists of machine-derived index terms plus WD-3 classification structure.

Document sets were numbered as follows:

1 = for use with the WD-2 classification structure

2 = for use with the Documentation Abstracts classification structure

3 = for use with the WD-3 classification structure.

Hence, a treatment code of A3 indicated a judgment set containing the full text of those documents selected for use with the WD-3 classification structure, and so forth. The experimental design is shown in Figure 32.

The design was severely limited by the number of subjects available. For reliability, it was desirable to have 10 judges in each set; however, a single judge could not be used on the same basic material more than once without danger of carry-over effects. This limited us to the 12 judgment sets shown. The design was guided by the necessity of making certain comparisons between surrogate types and the full text and the fact that some comparisons between different surrogate types were of only limited utility to the study. Each surrogate type was compared against at least two full text sets.

Notice in Figure 32 that full-text set No. 1 was judged twice, as A1 and A1', to balance out the design and to get an estimate of the reliability of the judgments of the full-text, etc.

Subject Group	Session I	II	III	N
1	D1	A2	B3	10
2	C2	E3	A1	10
3	A1'	C3	D2	8*
4	E2	A3	B1	9

*One subject made D2 judgments and did not complete the task.

Figure 32. A Balanced Design of the Experimental Conditions

a. Subjects

The experimental subjects were UCLA students in the Graduate School of Library Service. Personal information supplied by these subjects indicated that the majority were in their first semester of graduate study and most had had little or no experience with classification. The subjects were not randomly selected; they constituted the entire group who volunteered for the experiment except for the one subject who did not complete all three tasks. Four subjects, who could not attend the regular sessions, were given the tasks at SDC at a later time. Subjects were randomly assigned to subject groups, and were compensated for their time at a rate of $2.50 per hour.

b. Instructions

The subjects were given a one-hour instruction session during which the general instructions were read verbatim (Figure 33). Additional instructions are attached to the rating forms (Figure 34). These instructions were substantially the same for each rating set except for minor wording for each type of material to be rated. The remainder of the instruction session was spent on answering questions about the task and filling out a personal information form (Figure 35).

In the testing sessions, the subjects were provided with a dictionary of acronyms and obscure words (Figure 36) that occurred in the index lists, in addition to their test materials. During these sessions many subjects

UT Intr.

General Instructions and Orientation

Those of us who are conducting this study are employed as researchers by the System Development Corporation of Santa Monica, which is a non-profit corporation specializing in the design and development of large man-machine data processing systems.

One such class of systems we are interested in is library systems, and the present study you are participating in is concerned with one aspect of the library problem. I will now try to give you a brief sketch of the nature of this problem.

We are all aware of the tremendous increase in the number of scientific and technical publications per year. This increase, sometimes called the information explosion, is responsible for a correspondingly large increase in the work load of library workers. Particularly difficult is the indexing, classification, and document retrieval tasks. Our area of research deals with the area of machine-aided indexing and classification. We are trying to reduce the work involved in indexing and classifying documents by introducing machine-aided methods of both indexing and classification. However, such methods will be of little use if the results of a machine-aided indexing and classification system are of no use to the user. Any machine-aided system must produce an output that is as easily interpretable as the currently available systems.

One of the tools we are using in our research is a rating procedure known as paired comparisons. In this procedure you will be asked to rate (scale) your own personal judgment of the content of several technical articles and some representations (surrogates) of these articles. The surrogates consist of lists of index terms either man or machine produced (you won't know which), or such lists with added classification data. All in all there will be three judging tasks: one involving the full text of 12 articles, one involving index terms of 12 different articles, and the final task involving index terms and classifications of another 12 articles. You will be given the three tasks in different orders, that is, some of you will judge the articles first, and some the surrogates first, etc.

For the full articles, you will be given one week to read the articles at your leisure. For the surrogates, you will be given about one hour to familiarize yourselves with them.

The rating procedure is quite simple. On the second page of your rating form there are three columns of pairs of numbers. These are the numbers of the articles. You are to take the pairs of articles in the order that they appear in the columns and look them over to refresh your minds as to their content. Then you will make a numerical estimate of their apparent similarity using the scale on page one of the rating booklet. Do not worry about the exact meaning of the scale items; they are placed there as an aid to you in using the scale, but it is understood that each of you will adopt his own personal interpretation of the scale. All we ask of you is that you attempt to use the scale in as consistent a manner as you can. If you become tired, please feel free to take a break and leave the room. You will be given ample time to complete the task.

Figure 33. General Instructions and Orientation

However, we ask that you do not discuss any part of this task with others until the experiment is completed.

Remember this scaling task reflects your personal perceptions of the similarity of the documents; therefore, there is no right or wrong answer. You will not be scored in any such sense. Your judgments will be used in a subsequent mathematical analysis of the various indexing and classification systems.

If you desire, a report of the results of the analysis will be sent to you upon the completion of the study. If you wish such a report please indicate so in the place provided on your personal history sheet.

Also, although there is nothing in this task that reflects in any way upon you as individuals or as students, all responses will be kept anonymous according to the rules of the American Psychological Association. To aid this, you have all been assigned subject identification numbers; please place these numbers and only these numbers on each sheet of paper given you, including all pages of the rating booklets. Accuracy in the use of these identification numbers is extremely important, as is your care and attention to the rating task. Please check all your work carefully.

Are there any questions?

Figure 33--Concluded

Subject Identification Number UT 1-a

Document Similarity Rating Form

In this task you are asked to judge the similarity of the contents of pairs of documents. The document numbers are presented in pairs followed by a blank space. For each pair you are to make the best possible estimate of their similarity from the given information and indicate this judgment by selecting the statement below that comes closest to describing your judgment and placing its number in the blank opposite the pair being judged.

"To me, the subject contents of the two documents would most likely:

1. be almost completely similar."
2. be highly similar."
3. be quite similar."
4. be slightly more similar than different."
5. be about equally similar and different."
6. be slightly more different than similar."
7. be quite different."
8. be highly different."
9. be almost completely different."

About making judgments:

1. There is absolutely no basis in this experiment for considering any judgment you might wish to make as more or less "right" or "wrong." We desire your immediate, independent judgment, without consulting aids such as authority lists and without unduly extended analysis of the situation.

2. All document numbers occur again and again in different combinations in this exhaustive method of paired comparisons. The judgment task can become quite onerous, but we know of no other way to extract the needed detail of data. Accordingly, we depend on you to pace yourself as you see fit. If you notice your attention wandering or an inability to focus any longer on the task, please take a break and wait until you are able to return with fresh concentration.

3. Be sure to place your subject identification number on page 2 of the rating booklet in the upper left-hand corner. Place the list description number (A1, A2 ... E3) in the upper right-hand corner. This number is in the upper right-hand corner of the envelopes containing your materials and on each index term page (the list numbers are not on the full text document reproductions only on the envelope containing them).

Figure 34. Document Similarity Rating Form

Subject Identification Number	-2-	(List Designation)
7 - 12 _____	6 - 12 _____	1 - 7 _____
2 - 5 _____	5 - 7 _____	2 - 3 _____
1 - 6 _____	1 - 2 _____	5 - 10 _____
2 - 9 _____	5 - 9 _____	7 - 8 _____
10 - 11 _____	7 - 10 _____	6 - 11 _____
4 - 7 _____	4 - 11 _____	9 - 12 _____
5 - 8 _____	1 - 10 _____	1 - 9 _____
6 - 9 _____	3 - 7 _____	4 - 12 _____
1 - 4 _____	8 - 9 _____	8 - 10 _____
3 - 10 _____	5 - 12 _____	3 - 4 _____
3 - 11 _____	3 - 8 _____	5 - 6 _____
8 - 12 _____	2 - 12 _____	2 - 8 _____
4 - 6 _____	4 - 5 _____	1 - 11 _____
3 - 9 _____	8 - 11 _____	4 - 9 _____
1 - 12 _____	6 - 8 _____	8 - 12 _____
7 - 11 _____	4 - 10 _____	3 - 11 _____
2 - 4 _____	1 - 3 _____	3 - 10 _____
9 - 11 _____	2 - 6 _____	1 - 4 _____
1 - 5 _____	7 - 9 _____	6 - 9 _____
6 - 7 _____	10 - 12 _____	5 - 8 _____
5 - 10 _____	11 - 12 _____	4 - 7 _____
2 - 11 _____	5 - 11 _____	10 - 11 _____
1 - 8 _____	2 - 10 _____	2 - 9 _____
9 - 10 _____	3 - 5 _____	1 - 6 _____
5 - 6 _____	2 - 7 _____	2 - 5 _____
4 - 8 _____	3 - 12 _____	7 - 12 _____

Figure 34--Concluded

Subject Identification Number

Personal Information Form

1. Name _____ Age ____ Sex ____

2. Home Address _____ Phone _____

3. University _____

4. Status (grad., faculty, etc.) _____

5. Brief Description of Education--please note the major areas you have studied both as an undergraduate and graduate student, and your degree objective.

6. Work Experience--please list your major jobs, not part time or summer work. If in the library field, list type of duties and length of time.

Figure 35. Personal Information Form

Acronym	Definition
AEMIS	Medical IR system
AHU film	A type of microfilm sheet film
API	American Petroleum Institute
ASM	American Society for Metals
ASTIA, ASTIA thesaurus	Armed Services Technical Information Agency (predecessor to DDC)
CDCR	Center for Documentation and Communications Research
COBOL	A programming language
CONDEX	Concept indexing
COSATI	Committee on Scientific and Technical Information
DDC, DDC thesauri	Defense Documentation Center
DOD	Department of Defense
EJC, EJC thesaurus	Engineers' Joint Council
El-Nikkor	A camera lens
FORTRAN	Programming language
INFOL	An information language and index scheme
KWIC	Key word in context
KWOC	Key word out of context
LEX, project LEX	DoD project to develop common indexing vocabulary
Lodestar	Microfilm cartridge reader-printer
MEDLARS	Medical information retrieval system
MESH	Medical subject heading index
MHRST	Medical and health related sciences thesauri
microfiche	sheet microfilm
NMA	National Manufacturers' Association
NUCMC	National Union Catalog of Manuscript Collections
PAS	Personalized alerting service
RADC	Rome Air Development Center
RHD	Random House Dictionary
SYNTRAN	An indexing, abstracting and retrieval program
TEXT-90	An automated document preparation program
TEXTCON	A program for converting text into a better form for computers

Figure 36. Dictionary of Acronyms

had questions about the index terms and the interpretation of the tree/graph classifications. Questions of this nature were answered in general terms only, to avoid unduly influencing the subjects' rating. We feel that the manifest uncertainty concerning interpretation of the classification information had a considerable effect on the result. However, it is impossible to estimate the size of this effect.

c. Rating Procedure

The rating scale was presented during the instruction period and was reproduced on each of the rating forms (Figure 34). During the testing session, the subjects rated each pair of documents or surrogates on a 9-point scale of dissimilarity, one pair at a time. This resulted $(n^2-n)/2$ or $\frac{12^2-12}{2} = 66$ judgments. The first 12 judgments were arranged so that each article appeared at least one time; these 12 judgments were repeated at the end, thus bringing the total number of judgments made by each judge to 66 + 12 or 78. The repetition allowed the judges a 'warm-up' and some check on rater reliability. However, many subjects noted the repeated items and were then instructed to make the judgments again without referring to their earlier efforts. The first 12 judgments are used only for reliability checks and a points-of-view analysis. The fact that subjects noticed the repeats is of little consequence, since the major reason for the repeats was to allow for some warm-up to take place in each session and to assure that all 12 documents were referred to at least once before the major judging task.

For surrogate material the subjects were given the rating materials at the start of the judging session. For full-text material they were given the material one week before the task, with instructions to spend about six hours reading and familiarizing themselves with the material. They were allowed to make any notes they wished on the materials and the majority of the group did so.

On the whole, cooperation of the subjects was excellent and they appeared to have taken the task quite seriously and devoted good effort to the reading and judging of all materials.

3. DATA ANALYSIS

For the utility experiments, two separate but related data analyses were performed. The first was the points-of-view analysis designed to insure that no rater had a deviant approach to the rating task, and all judgments in a set could be combined. The second, or multidimensional scaling analysis, was designed to determine the number of aspects in the document or surrogate, such as the subject matter, difficulty level, writing style, etc., that contribute to the similarity judgments.

a. Points-of-View Analysis

The points-of-view analysis was adapted from a FORTRAN II coding originally done at the University of Southern California. This analysis follows exactly the procedures outlined in Tucker and Messick (op. cit.),

the SDC modifications being restricted to a recoding in FORTRAN IV (initially for the 7094 IBSYS operation and then for OS/360-65 operation).

In this study, a points-of-view analysis was performed on each judgment set. There were twelve such sets (see Figure 32), or unique combination of judges and stimuli. Each set provided an N x 78 matrix in which N was the number of judges and 78 was the number of judgments each rater made.

In all cases, that is, in all twelve sets of data, the analysis revealed the presence of only one dominant point of view. The largest root of the matrix accounted for over 90 percent of the trace. As a result, it was possible to combine the individual rater's judgments and to form a consensus judgment for each judgment set. The set consensus was computed by making a linear combination of judgments from each judge, using as weights the judge's score on the first factor of the points-of-view analysis. Thus, each judge's contribution to the consensus was in proportion to his 'distance' from the origin of the 'persons' space. This procedure, except for a normalizing factor, is equivalent to taking a weighted mean of the judgments as the consensus. In this study, the points-of-view procedure was used largely as a matter of convenience; the programs were already set up to work that way from previous research efforts, and the method was known to be superior to taking a simple average as the consensus, although an inspection of

the factor scores indicates that the weighted average is very little different from a simple average.

At the completion of the points-of-view analysis, each of the 12 experimental treatments had been reduced to a single vector of 78 judgments, one such vector for each set. Each set of vectors was rearranged by the program into a 12 x 12 document matrix of similarity judgments. A cell value in the matrix contained the consensus rating of the similarity of the pair of documents. In the process of forming the matrix, the first 12 judgments were deleted since these were repeated later on in the task. Twelve such matrices were formed--one for each experimental treatment. The matrices were formatted for direct input to the multidimensional scaling program.

b. Multidimensional Scaling Analysis

The multidimensional scaling program was recoded from a FORTRAN II version supplied by the authors (Torgerson and Meuser, opc. cit.) into FORTRAN IV, again first for the 7094 and then for the 360 computer. Both programs performed extremely well on author-supplied test problems. However, the 360 version of the multidimensional scaling program, for unexplained reasons, took three to four times the running time of the 7094 versions. This rather unexpected turn of events caused us to modify the normal procedure in using the multidimensional scaling program. Usually the program solves for the best-fitting space in successively

lower dimensions, starting with nine and iterating down to one. The output consists of a matrix of projections of items on each dimension, rotated to the principal axis position. This is a very lengthy procedure; the expected run times greatly exceed the time available to us for a single run. Therefore, a single solution in the highest dimensional space, nine, was used as the only solution in this experiment. Experience has indicated that the first 2 or 3 of the principal axis dimensions extracted change very little in the iterative process, and that, for 12 stimulus objects, the criterion of fit would be reached at about 5 or 6 dimensions, but the criterion of being able to relate dimensions obtained under different experimental conditions would apply to at most the first 3 dimensions. The time consideration was even more important than the not-inconsiderable cost. To follow the complete iterative procedure would have required at least three months, given the operating constraints now in existence with our new 360 system. The possible variation in results is very slight.

Only the first two dimensions extracted showed any positive relation over all relatable experimental conditions, so further analysis was restricted to these two dimensions.

4. SUMMARY OF RESULTS AND CONCLUSIONS

These experiments were designed to measure the degree of similarity between judgments made using the different document representations described in paragraph 2, Purpose and Methodology of the Utility Study. Since there were five different document representations, the number

of possible paired combinations (five objects taken two at a time) was equal to ten; that is, ten different comparisons among these five representations were possible. There were also 3 different document sets, and in a completely balanced design 30 comparisons could be made requiring 15 independent rating experiments. However, not all comparisons were of equal theoretic interest, and so only 11 and 1 replication (A1') were selected for detailed study. These 12 judgment sets are listed as the column and row headings in Figure 37. Note that the rows are divided to provide for the two dimensions (I and II) derived from the multi-dimensional scaling analysis. A total of 17 pairs of comparisons were made and are recorded in Figure 37. These comparisons indicate the degree of similarity or congruence between the configurations derived from judgments of different representations of the same documents.

Several different indexes of congruence have been suggested in the literature; they all share one common failing--none of them has known sampling properties. Therefore, no statements of 'statistical significance' can be made. The index used in this study is one suggested by Tucker (cited in Harmon, p. 257). It is essentially a product-moment type of index, but it is most definitely not a correlation coefficient. The formula is:

$$\lambda_{pq} = \frac{\sum_j {}_1a_{jp} \cdot {}_2a_{jq}}{\left[\left(\sum_j {}_1a^2_{jp}\right)\left(\sum_j {}_2a^2_{jq}\right)\right]^{1/2}}$$

Where $[{}_1a]$ and $[{}_2a]$ are the matrices of projection obtained under conditions 1 and 2.

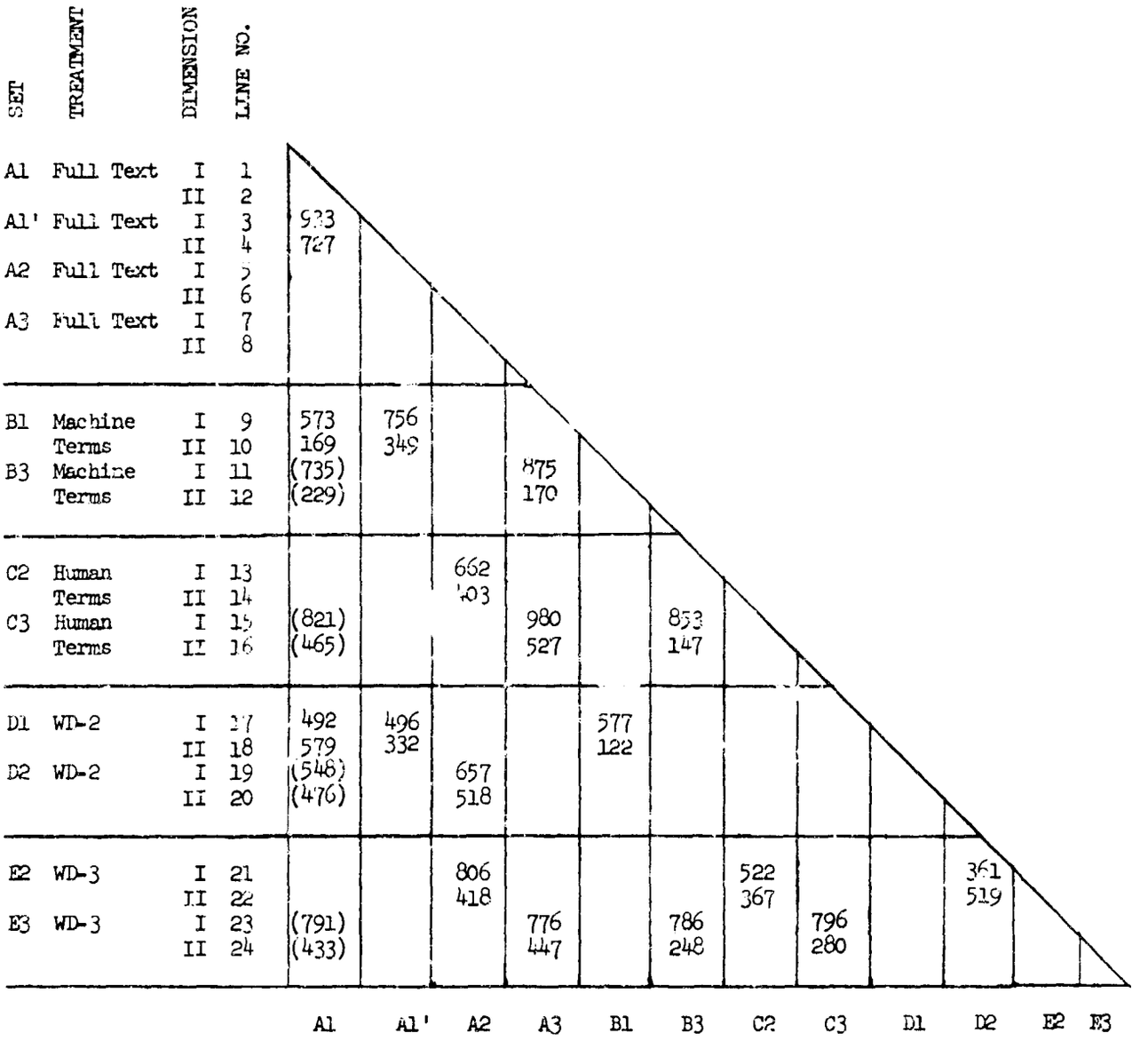

Figure 37. Degree of Similarity between Judgments of Different Representations of the Same Documents

As in a correlation, a value of 1.0 indicates perfect agreement and 0.0 indicates no agreement; since the sign of a projection is arbitrary, no special meaning can be attached to negative indices in the tabulated result.

In Figure 37, the index for the first dimension is placed directly above the index for the second dimension. Indices to the left of the double vertical line are those of greatest interest. Only the lower triangular matrix of indices is displayed, since the full matrix is symmetric around the diagonal.

Interpretation of Results

The points-of-view analysis, as was noted, produced no surprising results; therefore, that analysis can be viewed as simply a stage in the data processing without further comment.

The results of the multidimensional scaling analysis are summarized in Figure 37. In terms of this experiment, these indices are the best available summary of the results. Cross plots were made of dimensions I versus II for each judgment set. These plots were compared visually in the same combinations as indicated in Figure 37. However, only subjective estimates of congruence are possible by such a comparison, but these subjective estimates are accurately reflected in the indices presented (the visual comparisons were made before computing the indices). In the absence of a known sampling distribution of the congruence index,

certain ad hoc conventions were adopted. These at least follow accepted practice. An index of below .4 is assumed to indicate at best a trivial relationship; an index above .9 indicates a good relationship; and the points in between are interpolated along this scale.

Certain average indices were computed for convenience in intepreting Figure 37. Only indices shown to the left of the double vertical line were used in computing these averages. They are shown in the figure in parentheses, in the lower left corner of the block (single horizontal lines) from which they were derived. These indices are derived from on the order of 600 judgments per set and should be rather stable.

First, notice that the indices between sets* A1 and A1' are .933 and .727 (lines 3 and 4). This indicates that, for our sample of subjects, at least the first two dimensions (based on the information-rich full text) are reliable and replicable over different groups. Next, in general, the surrogates do not provide much information about the second dimension; the highest second dimension index is only .579 (line 18) for the D1 condition. However, that condition is one of the few that was replicated by comparing it against two full-text configurations, and the replication index is only .332, indicating that the degree of relationship is not very high.

*The labeling of these sets is fully described in paragraph 2.

Taken over all, the human-derived index terms contained the most information about the first dimension, with an average index of .821 (line 15). The machine terms did only slightly poorer with an average of .735 (line 11). The human terms might have a slightly greater edge in providing more second dimension information, an average of .465 (line 16) versus .229 (line 12). Adding the classification to the machine terms had an unpredicted effect--the WD-2 classification apparently depressed performance, reducing the average index for the first dimension to .548 (line 19), while adding the WD-3 classification improved things slightly, increasing the index to .791 (line 23). However, both classifications did add some second dimension information (lines 20 and 24), which was almost totally lacking in the machine index terms along line 12.

This result is somewhat hard to explain, since the judges had at least as much to judge on with the added classification information as with just the machine index terms. The decrement in performance can possibly be accounted for by the expressed difficulty of the subjects in interpreting the classification trees. The fact remains that subjects did use the classification data; if they had simply ignored the trees, one would expect no difference between conditions B (machine terms), D (WD-2), and E (WD-3). However, clearly, WD-2 was worse on Dimension I than either WD-3 or just the machine terms (lines 19 versus 11 and 23). WD-2 was perhaps slightly better than just machine terms on Dimension II as was WD-3. Further, those judges that used classification data were

relatively consistent among themselves, as can be inferred from the lack of secondary points of view in the points-of-view analysis. The most likely interpretation of the result is that the physical layout of the WD-2 trees led judges to overweight the fine distinctions between documents, represented by the 'leaf' end of the tree, simply because there were more of them. The physical form of the WD-3 trees did not mislead the judges as much; 'leaf' end clusters tend to go 'higher' in the tree than for WD-2.

To rephrase this, it appears that the WD-2 classification led the judges to consider the intracluster distances as being more salient than the intercluster distances. Multidimensional scaling has the property that, if clusters exist in the data, the analysis tends to disregard intracluster distances, treating the cluster like a point. Therefore, the WD-2 classification led the judges to consider a part of the information that would not be expected to show up in a multi-dimensional scaling analysis.

To check on this interpretation, a multidimensional scale analysis was performed on the whole 52-document set, using as distances the node-height measure described earlier. As expected, node heights that use less than five were mapped into multidimensional scale distances of zero for both the WD-2 and WD-3 classifications. The clusters of node height greater than five were found in the cross plots of the first two obtained dimensions. However, the WD-3 distances were, in general, greater than for WD-2, so that many more clusters were displayed in the first two dimensions, i.e., there was more intercluster information.

However, the surmise must remain just that, until more information is known about human classification performance in general.

WD-3 also fares better than WD-2 in other ways. Document set 3 was derived from WD-3 clusters of the whole 52-document collection. A general comparison shows that almost every index involving document set 3 was higher than other comparable conditions. Note the rows and columns labeled A3, B3, C3, and E3.

An analysis of variance would be inappropriate for these data; however, it is clear by inspection that there is a consistent 'Document Set Effect' in favor of Set 3. This is explainable if the WD-3 clusters are 'more distinctive' than the others, and, hence, documents and surrogates selected from WD-3 clusters are easier to distinguish.

Finally, the WD-3 classification data are substantially the same as the human index terms, on both dimensions I and II. The average indices for both dimensions are .643 (average of line 23 and 24) versus .611 (average of line 15 and 16) for human terms and WD-3 with first-dimension average indices of .821 (line 15) versus .791 (line 23). Based upon these data, it seems reasonable to assume that, as judges become more experienced in interpreting tree diagrams, the use of this form of automated classification would provide more information relative to the full text than would subject heading index terms alone.

At least, the relatively inexpensive indexing and classification represented by WD-3 is very nearly as informative to our class of judges as the much more costly human-derived index terms of condition C. Further, the machine terms alone (condition B) are fairly good relative to these same human terms.

SECTION VII

INTERPRETATIONS AND RECOMMENDATIONS

In preceding years, System Development Corporation, under contract with Rome Air Development Center, has developed a set of computer programs for the automated classification of documents. These programs, called ALCAPP (Automatic List Classification and Profile Production), were coded for use with the AN/FSQ-32 computer. In contrast with most other automatic document classification procedures, the ALCAPP programs are designed to be economical when used with large data bases, for computer time increases as a direct function of the number of items to be classified. The programming system consists of three parts: the data base generator, the hierarchical-grouping program, and the iterative cluster-finding program.

The current research project has as its purposes:

(1) To recode all three programs for use with RADC's GE 635 computer;

(2) To investigate the statistical reliabilities of the hierarchical-grouping program under a variety of conditions;

(3) To investigate the statistical reliabilities of the iterative cluster-finding program;

(4) To investigate the utility of the machine-produced classification hierarchies for predicting document content.

The preceding sections of this report describe in detail the experiments that were performed and the results that were obtained. In this

concluding section, we review and interpret the results and state our conclusions and recommendations.

1. RECODING THE PROGRAMS

All of the programs were rewritten in JOVIAL for compilation on the GE 635 computer. A detailed description of the programs and their flow charts is available in the Appendix to this report.

2. HIERARCHICAL-CLUSTERING PROGRAMS

The aim of this set of experiments is to compare the classification structures that are the result, or output, of the hierarchical-clustering program when various input conditions are manipulated. The conditions that were systematically varied and tested are:

(1) The classification algorithm;

(2) The type of indexing;

(3) The depth of indexing;

(4) The order in which the documents are processed.

The following paragraphs report the results in more detail, but in essence it can be stated that the output of the hierarchical-clustering program is sensitive to variations in the first three variables and relatively insensitive to order effect.

a. Interpreting the Effect of the Classification Algorithm

Two different classification procedures were compared, and it was determined that differences in the computer program result in document clusters that

are only moderately similar. While some clustering differences were expected, the classification structures were less similar to each other than anticipated.

These findings statistically support the view that, although all automatic classification techniques cluster documents on the basis of word similarity, the resulting classification structures may differ significantly from each other, depending upon the logic of the classification algorithm. Just as manual classification schemes differ from each other, so do mathematically derived classification systems. They are not the same, and the utility of each system must be evaluated separately.

As an outcome of this experiment, it can be stated that the WD-3 algorithm appears to be slightly more stable under a variety of input conditions than is the WD-2 algorithm.

Now that the structure and the statistical properties of both algorithms are known, it would seem advantageous to study methods of combining both logics--and indeed other logics as well--to develop a classification algorithm that would be more satisfactory than either one separately.

b. Interpreting the Effect of the Type of Indexing

The documents used in the hierarchical classification program were indexed by skilled librarians and by machine-aided techniques. The librarians prepared index lists of multiple-word concept terms while the computer derived individual key word index terms. The classification structures based upon both types of indexing were compared for similarity.

The results of this experiment are very significant, for they show that the automatically derived classification structures, based upon the same set of documents, will differ significantly, depending on whether the type of indexing used is key word or concept. The experiment demonstrates the need for a consistent vocabulary in classifying documents.

Although these findings are based upon machine-derived classification systems, the statistical significance of the results cautions against mixing concept and key-word indexing in any document storage and retrieval system.

c. Interpreting the Effect of the Depth of Indexing

A series of experiments were designed to investigate whether differences in indexing depth would result in differences in the classification structure. Lists of 6, 15, and 30 index terms were derived from the same document, and these lists were processed separately into classification structures, which were then compared for similarity.

It is concluded from the results that the number of index terms on the lists being processed can significantly affect the arrangement and clustering of items in an automatically derived classification hierarchy. Furthermore, this relationship holds true for both the WD-2 and WD-3 programs and for both key-word indexing and concept indexing. There is, however, an interesting difference based upon the type of indexing.

The longer the list of key words, the more stable is the classification structure. For human indexing, this trend is reversed and the classification structure is most stable when derived from lists containing relatively few multiple-word index terms.

A reasonable interpretation that can account for these results is that a fairly large number of key words are needed to make the index list an adequate (and thus stable) representation of the document. This is not true when using concept indexing. A relatively small number of concepts can adequately describe the subject matter of a document. If a larger number is used, extraneous concepts are included, and classifications structures derived from these longer lists are subject to chance fluctuations and are thus less reliable.

These experimental results are consistent with the previous findings that concept and key-word indexing should not be mixed. These findings are also significant in themselves, for they indicate that, certainly for machine derived classification structures and probably in general, there is an optimal number of index terms that makes for the most stable document surrogate, and this number differs, depending on whether concept indexing or key-word indexing is used.

d. Interpreting the Effects of the Order in which Documents Are Classified

In building a classification system, the original documents to be classified tend to exert a greater influence on the resulting structure than do later documents, which are then simply fitted into the existing structure. This statement seems to hold true for all classification systems, be they manually or machine derived. However, because of the manner in which documents are paired, the effect may be even greater in automatic classification procedures.

To investigate the effect that the order of document input may have, three different arrangements were used and the resulting classification structures compared. The results indicate that, while there is some variation in the final classification structure, processing order is not a very significant factor. It is also evident that the WD-3 algorithm is less sensitive to this variable than is WD-2. The use of concept indexing will tend to further increase the reliability of the classification structure.

3. ITERATIVE CLUSTER-FINDING PROGRAMS

In using these programs it is necessary to decide first on a reasonable number of categories and to arbitrarily assign a few documents to each of these initial categories. Then, by a series of iterations, the program will divide the entire document collection into groups.

Three questions were asked about the operation of this program:

(1) How dependent is the final cluster arrangement on the initial arbitrary assignment of documents?

(2) How stable are the clusters as new documents are added to the collection?

(3) How homogeneous and reasonable are the clusters?

The experimental results on which the answers to these questions are based are described in detail in Section V of this report. The overall conclusions and recommendations are that the iterative cluster-finding program is sensitive to changes in the initial cluster configuration, and that, therefore, in a practical application the initial documents should be selectively rather than arbitrarily assigned. By seeding the clusters with selected documents, we will obtain final categories that probably are more reasonable and homogeneous. The classification categories are stable, and additional documents can be added to the collection without causing any major shifts, provided that these new documents do not constitute more than ten percent of the original collection.

Finally, it is our conclusion that the automatically derived classification structure of a document collection constitutes a good initial organization of the material, but that this organization can be improved and made more meaningful if it is reviewed and modified by trained personnel.

4. THE USE OF MACHINE-PRODUCED CLASSIFICATION HIERARCHIES FOR PREDICTING DOCUMENT CONTENT

This set of experiments was aimed at determining whether knowing the classification category in which a document has been placed will provide additional useful information for judging the content of that document and therefore its possible relevance to a need for information. Essentially, the experimental design was based upon making judgments on how similar various document representations were as compared with the full document. We were particularly interested in knowing whether a document representation consisting of index terms and classification data was superior to a document representation of index terms alone.

A detailed description of these experiments and the results are available in Section VI.

In retrospect, it seems clear that the subjects needed more instruction and experience in using classification trees. Nevertheless, although they had difficulty in interpreting these trees, they did use the classification data in making their judgments. On the major variables most commonly used in judging document relevance, a knowledge of classification categories did not add much to the obtained scores. However, classification provided other information, as shown by the increased scores for the second dimension, and thus could improve the overall judgment of relevance.

5. FINAL RECOMMENDATIONS

The statistical properties of the Automatic List Classification and Profile programs have been clarified, and new knowledge has been gained about the strengths and weaknesses of these programs. To conclude that automated document classification is not perfect would be to make a true statement but one not based upon, or directly derived from, the results of the preceding experiments. It is a truism, as would be the statement that no existing library classification system is perfect, and it is just as meaningless.

Classification is a method of file organization, and it is needed in both traditional libraries and in automated document storage and retrieval systems.

Libraries employ skilled personnel to analyze the subject content of a document and to assign it a proper place in a logically organized structure. This is a time-consuming and expensive task, but it works reasonably well. However, in an automated system where every effort is being made to reduce search time and provide faster customer service, manual indexing and classification would be anachronistic. Why improve search time by microseconds when it takes weeks to put new documents into the file? Mechanization of the input procedure--the initial processing of the document and the organization of the file--is a necessity.

Many researchers have been working to achieve this goal. As is usually the case, in the beginning, great advances, ever breakthroughs, are made. But the consolidation of these gains and their application to practical systems is a long and painstaking task. The research reported in this study is a small but necessary step in making automated document classification a practical reality.

BIBLIOGRAPHY

Abelson, R. P. and Tukey, J. W. Efficient Conversion of Non-Metric Information into Metric Information. Proc. Am. Stat. Assoc., 1959. pp. 226-230.

Abelson, R. P. and Tukey, J. W. Efficient Utilization of Non-Numerical Information in Quantitative Analysis: General Theory and the Case of Simple Order. Anals. of Math. Stat., 1963, pp. 1347-1369.

Baker, F. B. The Ward Hierarchical Grouping Procedure: A Didactic. SDC document TM-2679, System Development Corp., Santa Monica, Calif., 1965.

Black, Donald V. (ed.). Proceedings of the 1966 ADI Annual Meeting. Adrianne Press, Woodland Hills, Calif., 1966.

Borko, H. Indexing and Classification. In: Automated Language Processing, H. Borko (ed.), John Wiley and Sons, Inc, New York, 1967.

Doyle, L. B. Breaking the Cost Barrier in Automatic Classification. SDC document SP-2516, System Development Corp., Santa Monica, Calif., 1966.

Harmon, H. H. Modern Factor Analysis, Second Edition. University of Chicago Press, Chicago, 1967.

Helm, C. E. A Multidimensional Ratio Scaling Analysis of Color Relations. Educational Testing Service, Princeton, New Jersey, 1959.

Helm, C. E. and Tucker, L. R. Individual Differences in the Structure of Color Perception. Amer. J. Psychol., Vol. 75, 1962, pp. 437-444.

Kruskal, J. B. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. Psychometrika, Vol. 29, No. 1, March 1964, pp. 1-27.

Kruskal, J. B. Nonmetric Multidimensional Scaling: A Numerical Method. Psychometrika, Vol. 29, No. 2, June 1964, pp. 115-129.

Shepard, R. N. Metric Structures in Ordinal Data. Jour. of Math. Psych., Vol. 3, 1965, pp. 287-315.

Shepard, R. N. The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. I. Psychometrika, Vol. 27, No. 2, June 1962, pp. 125-140.

Shepard, R. N. The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. II. Psychometrika, Vol. 27, No. 3, Sept. 1962, pp. 219-246.

Torgerson, W. S. and Meuser, G. Informal Notes on Torgerson and Meuser's IBM 7090 for Multidimensional Scaling. MITRE Corporation, Cambridge, Mass., 1962.

Tucker, L. R. and Messick, S. An Individual Differences Model for Multidimensional Scaling. Psychometrika, Vol. 28, No. 4, December 1963, pp. 333-367.

Ward, J. H., Jr. Use of a Decision Index in Assigning Air Force Personnel. Personnel Laboratory, Wright Air Development Center, WADC-TN-59-38. AD 214 600, April 1959.

Ward, J. H., Jr. and Hook, M. E. Application of a Hierarchical Grouping Procedure to a Problem of Grouping Profiles. Educational and Psychological Measurement, Vol. 23, 1963, pp. 69-92.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Systems Development Corporation
2500 Colorado Avenue
Santa Monica, California 90406

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP
N/A

3. REPORT TITLE

ON-LINE INFORMATION RETRIEVAL USING ASSOCIATIVE INDEXING

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Final Report

5. AUTHOR(S) (First name, middle initial, last name)
Harold Borko
Donald A. Blankenship
Robert C. Burket

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
May 1968	124	18

8a. CONTRACT OR GRANT NO
F30602-67-C-0077

b. PROJECT NO
4594

c.
459401

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
TM-(L)-3851

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
RADC-TR-68-100

10. DISTRIBUTION STATEMENT
This document has been approved for public release and sale; it distribution is unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Rome Air Development Center (EMIIH)
Griffiss Air Force Base, New York 13440

13. ABSTRACT

Experiments were performed to determine the feasibility of using ALCAPP as one form of on-line dialogue.

Assuming the ALCAPP (Automatic List Classification and Profile Production) system is in an on-line mode, investigations of those parameters which could affect its stability and reliability were conducted. Fifty-two full text documents were used to test how type of indexing, depth of indexing, the classification algorithm, the order of document presentation, and the homogeneity of the document collection would affect the hierarchical grouping programs of ALCAPP. Six hundred abstracts were used to study the effect on document clusters when more documents are added to the data base and the effect on the final cluster arrangement when the initial assignment of documents to clusters is arbitrary.

Results reveal that the only time significant differences in the classification of documents does not occur is when the order of document presentation is varied. Final clusters are significantly affected by the initial assignment of documents to clusters. The number of documents added to a data base allows stability of clusters only to a cutoff point which is some percentage of the original number of documents in the data base.

DD FORM 1473

UNCLASSIFIED

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
ASSOCIATIVE INDEXING CLUSTER-FINDING TECHNIQUES HIERARCHICAL GROUPING SCHEMES DOCUMENT CLASSIFICATION						